# Comprehensive Study of Germline Mutations and Double-Hit Events in Esophageal Squamous Cell Cancer

**Bing Zeng[1,2], Peide Huang[2], Peina Du[3], Xiaohui Sun[3], Xuanlin Huang[2], Xiaodong Fang[1,2,4*], Lin Li[2*]**

[1] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, 518083, China

[2] BGI-Shenzhen, Shenzhen, 518083, China

[3] BGI Genomics, BGI-Shenzhen, Shenzhen, 518083, China

[4] China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

*** Correspondence:**
Lin Li

Email: lilin_contact@163.com

Xiaodong Fang

Email: fangxd@bgi.com

**Keywords: Esophageal squamous cell cancer, cancer susceptibility gene, double-hit, germline mutation, pathogenicity**

## Abstract

Esophageal squamous cell cancer (ESCC) is the eighth most common cancer around the world. Several reports have focused on somatic mutations and common germline mutations in ESCC. However, the contributions of pathogenic germline alterations in cancer susceptibility genes (CSGs), highly frequently mutated CSGs, and pathogenically mutated CSG-related pathways in ESCC remain unclear. We obtained data on 571 ESCC cases from public databases and East Asian from the 1000 Genomes Project database and the China Metabolic Analytics Project database to characterize pathogenic mutations. We detected 157 mutations in 75 CSGs, accounting for 25.0% (143/571) of ESCC cases. Six genes had more than five mutations: *TP53* (n = 15 mutations), *GJB2* (n = 8), *BRCA2* (n = 6), *RECQL4* (n = 6), *MUTYH* (n = 6), and *PMS2* (n = 5). Our results identified significant differences in pathogenic germline mutations of *TP53*, *BRCA2*, and *RECQL4* between the ESCC and control cohorts. Moreover, we identified 84 double-hit events (16 germline/somatic double-hit events and 68 somatic/somatic double-hit events) occurring in 18 tumor suppressor genes from 83 patients. Patients who had ESCC with germline/somatic double-hit events were diagnosed at younger ages than patients with the somatic/somatic double-hit events, although the correlation was not significant. Fanconi anemia was the most enriched pathway of pathogenically mutated CSGs, and it appeared to be a primary pathway for ESCC predisposition. The results of this study identified the underlying roles that pathogenic germline mutations in CSGs play in ESCC pathogenesis; increased our awareness about the genetic basis of ESCC; and provided suggestions for using highly mutated CSGs and double-hit features in the early discovery, prevention, and genetic counseling of ESCC.

# 1 Introduction

Esophageal squamous cell cancer (ESCC) is one of the most common cancers in the world, and it is especially common in Asian countries, North America, and the eastern corridor of Africa (1). In China, there are approximately 478,000 new cases and 375,000 deaths related to ESCC each year (2). Many factors reportedly have relationships with ESCC; these include smoking, drinking, and dietary habits (3). However, the hereditary factors involved in ESCC remain unclear. Thus, understanding the genetic mutations and molecular events in ESCC might be pivotal to reduce the incidence and mortality rate of ESCC.

Enormous efforts have been taken to identify somatic alterations by whole-genome sequencing (WGS) or whole-exome sequencing (WES) (4,5), and several studies reveal the complex process of tumor development (6,7). Many common germline single-nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies (8–16). rs138478634, a *CYP26B1* low-frequency variant, was proved to be involved in ESCC development (14). In 2018, several pan-cancer studies focused on pathogenic germline mutations to explore hereditary factors in cancers; 871 rare cancer predisposition mutations and copy number variations (CNVs) were observed in 8% of 10,389 cases, and 7.6% of the 914 patients with pediatric cancers had tumors that harbored pathogenic mutations in cancer predisposition genes (17,18). Deng et al. (19) identified germline profiles in Chinese patients with ESCC and uncovered the association between genotype and environment interactions. Additionally, *BRCA2* was associated with ESCC risk in Chinese patients (20). Reflecting a critical part of cancer susceptibility, the two-hit hypothesis assumes that hereditary retinoblastoma involves double mutations and that one mutation is in germline DNA whereas nonhereditary retinoblastoma involves two somatic mutations (21). On the basis of these findings, double-hit events in some studies were used to identify cancer predisposition genes (22,23). These studies demonstrated the significance of pathogenic germline mutations and double-hit events in genetic testing and risk assessment for cancer.

To our knowledge, cancer predisposition genes and molecular events in ESCC remain poorly understood. Here, we identified pathogenic/likely pathogenic germline predisposition mutations and highly frequently mutated CSGs in a large ESCC cohort. We discovered significantly different pathogenic germline mutations of *TP53*, *BRCA2*, and *RECQL4* in ESCC cohorts, and we clarified the association between double-hit events and diagnosis age in patients with ESCC. In addition, we identified pathogenically mutated CSG-related pathways for ESCC to illuminate the mechanism affected by pathogenic mutations. Results of this study will improve genetic testing for relatives of patients with ESCC and facilitate implementation of organizational or institutional measures for ESCC prevention and surveillance.

# 2 Materials and Methods

## 2.1 Sample acquisition

We collected 592 ESCC samples from published studies and The Cancer Genome Atlas (a total of nine projects) (Supplementary Table 1), and we excluded poor-quality samples and hypermutant samples (4,5,24–29). Clinical information is listed in Supplementary Table 2. The WGS and WES data from the same studies came from distinct patient cases. The quality control analysis uncovered an average sequencing depth of 55×~161× for WES samples and 30×~65× for WGS samples (Supplementary Figure 1A), the 10× average coverages were more than 90% in most WES and WGS samples (Supplementary Figure 1B). Moreover, the relationship between 10× average coverages and average sequencing depths showed a positive correlation (Supplementary Figure 1C), suggesting that

the qualities of most samples were proofed. The mean depth of our data and the public databases we used as controls were able to provide enough variants to execute downstream analysis (30). The study protocol was reviewed by the institutional review board of the Beijing Genomics Institution.

### 2.2 Data processing and mutation calling

The fastq data from 571 samples (38 WGS samples, 533 WES samples) were trimmed and filtered using SOAPnuke (v1.5.6 with default parameters, except where -n 0.1 -l 11 -q 0.5 -G -T 1) (31). Data from ESCC-P006 was transformed from bam files using the GATK SamToFastq (v4.0.6.0 with default parameters) (32). The high-quality reads were aligned to the hg19 human reference genome with a Burrows-Wheeler Aligner (v0.7.17-r1194-dirty with default parameters, except where -o 1 -e 50 -m 100000 -i 15 -q 10 -a 600) (33). MarkDuplicates GATK (version as above with default parameters, except where -CREATE_INDEX true, -reportMemoryStats true, -VALIDATION_STRINGENCY SILENT) was used to mark duplicated reads. BaseRecalibrator (version as above with default parameters) and ApplyBQSR (version as above with default parameters, except where -create-output-bam-index true) were performed to base quality score recalibration (32). Germline variants were joint-called using GenotypeGVCFs (version as above with default parameters, except where -ignore-variants-starting-outside-interval true) after CombineGVCFs (version as above with default parameters) and annotated with the Variant Effect Predictor (VEP v98.3) (32,34). The calling germline variants of nine projects are shown in Supplementary Figure 1D. Samples with fewer than 80,000 variants were filtered out. Somatic variants were detected by GATK MuTect2 (version as above with default parameters except where -af-of-alleles-not-in-resource 0.0000025, -native-pair-hmm-threads 1, -add-output-vcf-command-line false), and Oncotator (v1.9.9.0) was used for annotation (32,35). Loss of heterozygosity (LOH) and other somatic CNVs (SCNVs) were detected with FACETS (v0.5.14) and Pathwork (v1.0) for 533 WES and 38 WGS samples, respectively (36,37).

### 2.3 CSG sets

We curated CSGs from published papers and the Catalogue of Somatic Mutations in Cancer (COSMIC) database (38); we included cancer predisposition genes from three papers (17,18,39) and genes with recorded germline associations in COSMIC (Supplementary Table 4). After we removed duplicated genes, the CSG set included 260 genes. CSGs were divided into three groups according to the literature (17,40–42); these groups were tumor suppressor genes (TSGs; n = 139), oncogenes (n = 36), and nonclassified genes (n = 85).

### 2.4 Pathogenicity evaluation

We first leveraged an in-house pathogenicity database to match germline variants; the rest of the germline variants were evaluated using InterVar (InterVar_20190327) as a supplemental method to find germline pathogenic/likely pathogenic mutations (43). Germline pathogenic or likely pathogenic variants are hereafter referred to as pathogenic mutations. The pathogenicity database included ClinVar, the Human Gene Mutation Database, mutations collected from papers, and mutations we assessed according to consensus guidelines by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (17,44–46). We filtered for pathogenic variants with an allele frequency of 0.5% or lower in the Genome Aggregation Database (gnomAD version v2.1) (47). Pathogenic mutations in 260 high-interest CSGs (Supplementary Table 6) were selected for analysis and were checked by Deep Variant (48); manual verification ruled out false-positive results. For somatic nonsilent variants, with the exception of frameshift, nonsense, and splice-site mutations, three silico tools—SIFT (49), Polyphen2_HDIV (50) and CADD (51) were

used to predict pathogenicity. If a variant was predicted as damaging in any two silico tools (SIFT: D, Polyphen2_HDIV: D/P, CADD score > 15), the variant was categorized as deleterious (39,52).

### 2.5 Identification of potential double-hit events

According to the two-hit hypothesis, potential double-hit events are identified after two or more hits have been found in the same CSG; in this study, we set rigorous standards for determining hits. Pathogenic germline mutations were considered hits. Effective somatic variations were defined as hits if they met the following requirements: frameshift, nonsense, splice-site mutations, or deleterious missense and in-frame variants and SCNVs that caused allele loss. Copy-neutral LOH, duplication LOH, homozygous deletion, and hemizygous deletion were assumed to be linked to allele loss and were termed allele loss SCNVs (53,54). Integrative Genomics Viewer software was used to examine the authenticity of biallelic events (55). For double-hit events comprised of germline hits and allele loss SCNVs, we calculated SNP average depths and variant allelic frequency in normal and tumor tissues of ESCC to further validate allele loss SCNV events. Samples with variant allele frequencies less than 0.5 in tumors were removed.

### 2.6 Statistical analyses

To evaluate the correlations of the clinical features and genetic events, we used the two-sided Student's t-test. We conducted the two-sided Fisher's exact test to assess the gene-based association analysis and pathway enrichment. We also performed a burden test to determine the exact relationships between pathogenic mutations in CSGs and ESCC (56); $p < 0.05$ was defined as statistically significant.

## 3 Results

### 3.1 Population characteristics

Overall, 469 of 571 patient cases were Asian (424 Chinese, 41 Vietnamese, one Canadian, one Brazilian, two without country information), 41 were Caucasian, 58 were Black or African American, and the rest were Brazilian without ethnicity information. The entire population consisted of 105 women, 465 men, and one patient without gender information. The average diagnosed age for 567 patients (the rest had no information) was 58.81 years (the minimum diagnosed age was 24 years, and the maximum diagnosed age was 93 years). Thirty-five patients had family histories of ESCC, and the average age of patients with ESCC with a family history (mean age and standard deviation [SD] was 56.80 [9.3] years; [range, 41–82 years]). This average was lower than the age of patients with ESCC without a family history (mean age [SD], 60.00 [8.2] years; [range, 36–78 years]; t-test $p = 0.059$; 95% CI, -6.511 to 0.121) (Supplementary Figure 2). The average survival for 399 patients (the rest had no information) was 879.8 days (minimum survival, 3 days; maximum survival, 2,580 days). In this study, 347 patients had a smoking history, and 215 patients had histories of alcoholism. With regard to disease grade, 334 patients had disease with pathological grade 2 or lower, and 86 patients had disease with pathological grade greater than 2; pathological grade information was missing for 151 patients. All patients were diagnosed with disease stages I (n = 72), stage II (n = 207), stage III (n = 203), and stage IV (n = 7); 82 patients were not assigned disease stages for this study (their information was lost).

[Insert Figure 1 here]

### 3.2 Pathogenic germline mutations in CSGs

Overall, 2,484 pathogenic germline mutations were identified, including 1,973 SNPs and 511 insertions or deletions (Supplementary Table 5). Each sample had an average of 4.4 pathogenic mutations. After filtration by CSGs, 157 pathogenic mutations (113 SNPs and 44 insertions or deletions) were discovered from 25.0% (143/571) of the population (Supplementary Figure 3). Although each sample had an average of 1.1 pathogenic mutation in CSGs, only 12 (2.10%) of the 571 patients harbored one or more pathogenic mutation in CSGs (Figure 1, Supplementary Table 6). The frequency of most mutations was rare in the gnomAD noncancer database and in the China Metabolic Analytics Project (ChinaMAP) database (47,57), indicating the sparsity of these deleterious mutations in the general population. As expected, most of the frequently mutated CSGs belonged to TSGs, and they were involved in biological processes, such as DNA repair.

In general, the CSGs detected more than five times were *TP53* (n = 15 mutations), *GJB2* (n = 8), *BRCA2* (n = 6), *RECQL4* (n = 6), *MUTYH* (n = 6), and *PMS2* (n = 5). *TP53* was the most frequently mutated CSG, with pathogenic germline mutations in 2.63% (15/571) of patients with ESCC (Figure 1, Supplementary Table 6 and Supplementary Figure 4); The result was the same as *TP53* pathogenic mutations in a study of osteosarcoma (39). In our study, 86.7% (13/15) of *TP53* mutations were nonsynonymous single-nucleotide variations. c.A1073T (rs773553186; in 0.35%, or 2/571) and c.C742T (rs121912851; in 0.18%, or 1/571) were recorded in the International Agency for Research on Cancer TP53 database (58). All *TP53* pathogenic mutations were found in Chinese patients, except c.A1073T (one each in a Chinese and a Caucasian patient) (Supplementary Figure 4). Three of the *TP53* mutations, c.C742T, c.C586T, and c.C817T, have been reported in osteosarcoma (39), and *TP53* c.C742T has also been identified in low-grade glioma (17) (Supplementary Figure 4). Pathogenic mutations in *GJB2* represented the second most frequently mutated CSGs (Figure 1); their detection rate was 1.40% (8/571). The c.235delC (rs80338943) mutation, a common pathogenic frameshift deletion mutation in East Asian (EAS) populations, has been detected in six Asian (Chinese) patients with ESCC (59). Because this mutation has not been detected in other populations, rs80338943 may be specific to Chinese or Asian populations.

Nonsynonymous single-nucleotide variations occupied no less than 50% of pathogenic germline mutations in *BRCA2*, *RECQL4*, and *MUTYH* (Supplementary Table 6). In the upstream region, we detected a pathogenic splice mutation, *BRCA2* c.-39-1_-39delGA (rs758732038), in a patient, and the mutation was reported in ClinVar as likely pathogenic (46). The mutation has also been reported in patients with breast cancer and medulloblastoma (60–62). *RECQL4* pathogenic mutations were only detected in Asian (Chinese) patients in our study, and *RECQL4* c.C2272T has been reported in ovarian cancer/Rothmund–Thomson syndrome. In our study, *MUTYH* c.C1178T (rs36053993) and c.C458T (rs762307622) were detected three times (0.53%, or 3/571) and two times (0.35%, or 2/571), respectively. rs36053993 only detected in Caucasian patients and rs762307622 only detected in Asian (Chinese) patients. From gnomAD, rs36053993 in a homozygous state was found in three non-Finnish Europeans; this mutation may have been caused by founder events (63,64). Pathogenic mutations in *PMS2* were detected five times in five patients in our study (0.88%), and c.2192_2196delAGTTA (rs63750695) was observed in only four patients, who were all African. The rs63750695 mutation has also been discovered in Lynch syndrome, colorectal cancer, and ovarian carcinoma (65–67); however, it was rare in noncancer gnomAD and ChinaMAP, for which frequencies were $1.15 \times 10^{-5}$ and 0, respectively (Figure 1). rs63750695 is possibly specific to African ethnicity in ESCC.

The total number of pathogenic germline mutations and the frequency of mutations were relatively lower in oncogenes and nonclassified genes compared with TSGs. *TSHR* and *MPL* were oncogenes

that were mutated in two patients with ESCC; other oncogenes occurred in just one patient. *SLC25A13* was one of the nonclassified genes with the most pathogenic mutations.

We also investigated our pathogenic germline mutations in a previous pan-cancer study (17). Nine mutations were spread over 22 samples with diverse cancers (Supplementary Table 9). *SLC25A13* c.852_855delCATA (n = 7), *GJB2* c.235delC (n = 7), and *PALB2* c.C2257T (n = 2) were the variants observed more than once across cancers. We detected multiple susceptibility loci (31/47), also identified in previous genome-wide association studies, in our patients with ESCC (Supplementary Table 10) (8–16). Of those genes with susceptibility loci, pathogenic mutations *PDE4D* c.T108A and *RUNX1* c.61+1delG were found in two patients separately (Supplementary Table 5). We also confirmed from the COSMIC database that 87.3% (137/157) of pathogenic mutations in CSGs had nonsilent somatic mutations in the same or a nearby (within five) amino acid position (Supplementary Table 6). Among 137 mutations, 107 mutations were observed in TSGs, representing 89.2% (107/120) of all mutations.

[Insert Table 1 here]

### 3.3 Pathogenic germline mutations frequency in ESCC cases versus controls

To reveal the relationships between highly frequent mutated CSGs and ESCC, we chose the Chinese patients to continue the study, to leverage the most population data and avoid any ethnicity-specific effect. We conducted gene-based association analyses by comparing various germline mutation data from individuals with ESCC versus a 1000 Genomes Project EAS population and versus a ChinaMAP population separately (57,68). We also conducted rare variant burden tests on the ESCC individuals and the 1000 Genomes Project EAS population (68). Through the same pathogenicity evaluation pipeline, pathogenic mutations were identified in two public database populations. Analysis of results identified significantly higher pathogenic mutations in Chinese patients with ESCC versus public population databases (including 1000 Genomes Project EAS and ChinaMAP data), as reflected by odd ratios (ORs) of pathogenic mutations in *TP53* from the Chinese ESCC populations compared with the 1000 Genomes Project EAS populations (OR = 4.26; 95% CI, 1.33-17.91; Fisher's exact test $p = 7.359 \times 10^{-3}$) and compared with the ChinaMAP populations(OR = 10.59; 95% CI, 5.21-20.45; Fisher's exact test $p = 1.851 \times 10^{-9}$); in *BRCA2* from the Chinese ESCC populations compared with the 1000 Genomes Project EAS populations (OR= infinity; 95% CI, 1.09-infinity; Fisher's exact test $p = 0.0197$) and compared with the ChinaMAP populations (OR = 2.68; 95% CI, 0.83-6.75; Fisher's exact test $p = 0.0489$); and in *RECQL4* from the Chinese ESCC populations compared with the 1000 Genomes Project EAS populations (OR = 7.21; 95% CI, 0.87-332.23l; Fisher's exact test $p = 0.0519$) and compared with the ChinaMAP populations (OR = 3.69; 95% CI, 1.27-8.81; Fisher's exact test $p = 0.0089$)(Table 1). Likewise, in the burden analyses (Table 1), the numbers of pathogenic mutations from *TP53* (14/424, or 3.30%; burden test $p = 3.050 \times 10^{-3}$), *BRCA2* (5/424, or 1.18%; burden test $p = 0.015$), and *RECQL4* (6/424, or 1.14%; burden test $p = 0.035$) in our Chinese ESCC cohort were higher than those observed in the 1000 Genomes Project EAS group.

[Insert Figure 2 here]

### 3.4 Potential double-hit events

To further survey the genetic predisposition of ESCC, we tried to identify potential double-hit events in ESCC. First, we identified 49,876 nonsilent mutations (Supplementary Table 3) in protein-coding regions from patients with ESCC. (We filtered the somatic mutations that overlapped with our own

panel of normal datasets and the Exome Aggregation Consortium database.) Then, by integrating pathogenic germline mutations and effective somatic mutations (Supplementary Table 8) or allele loss SCNVs, we found 84 potential double-hit events (Figure 2). To distinguish hits with germline mutations, the double-hit events were classified as germline/somatic double-hit events and somatic/somatic double-hit events. We identified 16 potential germline/somatic double-hit events (two germline mutations coupled with somatic mutations, and 14 germline mutations accompanied with allele loss SCNVs) (Figure 2, Supplementary Table 11 and Supplementary Figures 5, 6) in 16 patients with ESCC, and we identified 68 potential somatic/somatic double-hit events (three somatic mutations accompanied by allele loss SCNVs and 65 double somatic mutations) (Figure 2, Supplementary Table 12) in 67 cases. The likelihood of two or more somatic mutations happening on the same chromosome was very low (52,69,70). Therefore, we assumed that double somatic mutations were likely in the trans position. Briefly, 83 individuals with ESCC possessed potential double-hit events, representing 14.5% of the ESCC cohort (Figure 2). Notably, one patient had two somatic/somatic double-hit events in different genes.

*GJB2* and *TP53* were the top two CSGs that found germline/somatic double-hit events. Germline/somatic double-hit events were identified in eight CSGs, including *BRCA2*, *BRCA1*, *MUTYH*, *CDKN2A*, and *ATM*. The dominant type of germline/somatic double-hit events was a germline mutation accompanied by an allele loss SCNV. In the remaining, germline mutations were coupled with somatic mutations; these were only discovered in *TP53* and *BRCA1*, possibly because SCNVs are relatively abundant in tumors and cover large genome region. In the somatic/somatic double-hit events, the *TP53* gene had the highest frequency, and most of the remaining genes had one potential double-hit event. Double somatic mutation was the main type of somatic/somatic double-hit event (Supplementary Table 12).

[Insert Figure 3 here]

When we compared diagnosis ages of patients with different double-hit events, we found that patients with germline/somatic double-hit events (with pathogenic germline mutations) had younger diagnosis ages (mean age [SD], 54.6 [11.2] years; [range, 36–71 years]) compared with patients in the somatic/somatic double-hit events (without pathogenic germline mutations; mean age [SD], 60.6 [7.8] years; [range, 4–80 years]; t-test $p = 0.056$; 95% CI, -12.216 to 0.177) (Figure 3). The comparison was nonsignificant, maybe it's due to the limited number of samples with double-hit events in this comparison. However, the finding was consistent in the study by Knudson (21). Using the empirical cumulative distribution function (ecdf) to calculated the expression percentiles of TSGs in an ESCC-P006 cancer cohort, two patients with somatic/somatic double-hit events showed low expression: one in *TP53* (5.32%) and one in *PTEN* (6.38%) (Supplementary Figure 8) (17). Those results support the two-hit hypothesis and suggest that genetic screening in specific TSGs can detect patients with germline/somatic double-hit events earlier.

[Insert Figure 4 here]

### 3.5 Pathway enrichment

To obtain a more comprehensive understanding of pathogenic germline genetic mutations affecting pathways, Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses were performed for multiple gene lists. The Fanconi anemia (FA) pathway was the most significantly enriched in the analysis of 75 pathogenically mutated CSGs (Fisher's exact test $p = 6.634\times10^{-19}$) (Figure 4A, Supplementary Table 7). In addition, 1,226 pathogenic mutated genes and the genes involved in germline/somatic double-hit events were significantly enriched in this pathway. The top four

pathways for CSGs involved in somatic/somatic double-hit events versus for CSGs involved in germline/somatic double-hit events differed significantly (Supplementary Figures 7A-D).

In the tumor-suppressor network, the FA pathway functions to preserve genomic integrity by repairing DNA interstrand crosslinks, regulating cytokinesis, and mitigating replication stress (71,72). 33 ESCC samples carried pathogenic mutations in 13 CSGs included in the FA pathway (Figure 4B). The homologous recombination pathway and the mismatch repair pathway described in a previous ESCC project, and associated with cancer susceptibility, were found in our study (Supplementary Figure 7A) (19,73–75). Those pathways were also reported in pathway enrichments of ovarian cancer and osteosarcoma (39,76). We also interrogated the oncogenic signaling pathways upon which our mutated CSGs converged (77). The cell cycle pathway was the most enriched, followed by p53 pathway, the phosphatidylinositol 3´-kinase-Akt pathway, and the receptor tyrosine kinases-Ras pathway.

## 4 Discussion

We reported the profile of pathogenic germline mutations of a larger ESCC cohort comparing with previous studies (17,19). We found 157 pathogenic mutations in CSGs from 143 (25.0%) of 571 patients with ESCC and identified 84 double-hit events in 83 individuals (14.5%). Double-hit events were found in almost all projects in our study except ESCC-P008, which demonstrated that double-hit events are relatively common in ESCC. As far as we know, there was no report about pathogenic mutations in *GJB2*, *RECQL4*, *MUTYH* and *PMS2* in ESCC, however, they were discovered in our study. Overall, *TP53*, *GJB2*, *BRCA2*, *RECQL4*, *MUTYH*, and *PMS2* were highly frequently mutated CSGs. Significant pathways were identified for different CSGs with pathogenic mutations; the FA pathway appeared to be a primary pathway for cancer predisposition in ESCC. We showed that significantly more pathogenic mutations from *TP53*, *BRCA2* and *RECQL4* occurred in patients with ESCC than in control cohorts, which indicates that these three CSGs may play vital roles in ESCC. Interestingly, *TP53* and *RECQL4* have also been found significantly associated with osteosarcoma (39). The relationship with diagnosis age was not significant in our study, but double-hit events may be pivotal in ESCC carcinogenesis.

We found that *TP53* had the highest frequency of pathogenic germline mutations and the most double-hit events in CSGs. In our study, 80% (12/15) of germline mutations in *TP53* were located in the p53 domain, which functions in DNA binding. This domain contains four conserved regions that are enriched for somatic mutation hot spots and are essential for the function of the TP53 protein as a transcription factor (78,79). Six of the 12 mutations were discovered in conserved regions. Environmental factors and specific DNA sequences drive higher mutation rates, which may explain why p53 domain was a hot-spot region (80). Those pathogenic *TP53* mutations may disrupt the p53 transcriptional pathway, which would enhance tumor progression and metastatic potential (81). The US Food and Drug Administration had approved drugs against the pocket in p53 domain (82). These drugs provide treatment options to patients with tumors that have mutations in the p53 domain. Results of studies in other cancers contrast with our findings about *TP53*. In a renal cell carcinoma study, *FH*, instead of *TP53*, harbored the most double-hit events, and *BRCA1* harbored the most in a pan-cancer study(17,22). Previous studies have reported that most double-hit events with *TP53* involve a mutation accompanied by LOH (83,84). However, in our research, double somatic mutations were the dominant type of double-hit event. It was partially due to the lack of researches on *TP53* double somatic mutations before.

*BRCA2* and *RECQL4* harbored more pathogenic germline mutations in ESCC than in public population. *BRCA2* is known for its involvement in breast cancer and ovarian cancer via the homologous recombination pathway, which is essential for repairing damaged DNA (85,86). And studies have reported *BRCA2* mutations related to ESCC risk in Chinese and Turkmen populations (20,87,88). The double-hit events detected in *BRCA2* in our study were germline/somatic double-hit events; the germline mutations were accompanied by allele loss SCNVs. These results were distinct from those reported in pancreatic acinar-cell carcinomas (89). *RECQL4* is a TSG that encodes RECQL4 helicase, which is involved in DNA replication and DNA repair. Germline mutations in *RECQL4* can cause Rothmund-Thomson syndrome and sporadic breast cancer (90). Although the pathogenic mutations in our ESCC cohort and in the 1000 Genomes EAS group were not significantly different (Fisher's exact test $p = 0.0519$), the difference between them was also confirmed by analysis of the ChinaMAP cohort (Fisher's exact test $p = 0.0089$). Importantly, this is the first report, to our knowledge, that illustrates the role of pathogenic mutations in *RECQL4* in ESCC.

The PMS2 protein is a homolog of the PMS1 protein (91) and both of them are components of the mismatch repair system. Common polymorphisms of *PMS1* have been positively associated with ESCC in an African population (92). This finding, together with the connection between PMS1 and PMS2, suggests a possible relationship between *PMS2* and ESCC. Double-hit events of mismatch repair genes could result in Lynch syndrome, as described in several studies (70,93), but we did not detect double-hit events in *PMS2* in our ESCC cohort. A larger ESCC cohort study might uncover double-hit events in *PMS2*, which would strengthen our understanding about ESCC susceptibility.

The genetic variations in ESCC are complicated. Although not all ESCC samples carried pathogenic germline mutations in CSGs, the detection rate of pathogenic mutations was close to that found in osteosarcoma (39). Because numerous susceptibility loci reported in genome-wide association studies were found in this research, we acknowledge that pathogenic mutations and known susceptibility loci may inform a genetic basis of ESCC. Our findings of variants and genes shared between ESCC and other cancers suggests that common hereditary factors exist in pan-cancer. Given the interplay of common SNPs and pathogenic mutations reported in breast cancer and colorectal cancer, the interaction between susceptibility loci and pathogenic mutations in ESCC suggests a need for future exploration (94).

To better understand the genetic factors causing ESCC initiation and development, we confirmed the putative germline-somatic interplay by COSMIC proximity match. The results not only support the pathogenicity of those germline mutations but also imply a signal functional relevance between germline and somatic mutations (76). In addition, we identified potential double-hit events in 83 patients with ESCC; though the difference was not significant, the patients with germline/somatic double-hit events were more likely to be diagnosed at younger ages. It is possible that pathogenic mutations confer the earliest genetic hits to TSGs in cells, so a somatic hit alone would cause loss of function in TSGs (95). As a result of double-hit events, the cells generate malignancy. Furthermore, enriched pathways revealed the process of pathogenic mutations that affect ESCC tumorigenesis and development. In patients without pathogenic mutations or double-hit events, limited CSG sets, potential alternations in methylations of a promoter region, germline CNVs, and gene-environmental or gene-lifestyle interactions are possible explanations for ESCC development.

Despite our findings about the genetic characterization of and double-hit events in ESCC, we still acknowledge limitations to our study. The first is our inability to obtain detailed clinical information because of limited access to public databases. Second, merging different data, such as WGS and

WES, may induce biases in cohort-wide variant processing. Third, directly adopting variants from different sources may influence comparisons, because the different sources applied distinct platforms and variant detection pipelines. Fourth, our sample size was not large enough for statistical tests, especially for individual variants.

In sum, we report that approximately 25.0% of patients with ESCC harbored at least one pathogenic germline mutation in CSGs, and approximately 14.5% of ESCC cases could be explained by a two-hit hypothesis. Significantly enriched pathways also validated the significance of those pathogenic mutations. Myriad genome variations occur in patients; our findings represent, to our knowledge, the largest discovery of rare, germline predisposition mutations in ESCC so far. These results strengthen the understanding about genetic factors involved in ESCC and will help improve prevention, early detection, and risk management of ESCC for patients. We acknowledge the shortcomings in the analytical methods and the data sources used. Additional studies are needed to improve our observations and results.

## 5 Data availability statement

The raw data of this project can be found in the Sequence Read Archive hosted by the National Center for Biotechnology Information under accession numbers PRJNA315775, PRJNA399748, PRJNA401209, PRJNA230271, PRJNA317404, SRA112617 and SRP034680 and in the European Genome-Phenome Archive under accession number EGAD00001000845.The whole-exome sequencing data of esophageal squamous cell cancer samples from The Cancer Genome Atlas are available from the National Cancer Institute Genomic Data Commons (https://portal.gdc.cancer.gov/). All relevant datasets for this study are available from the authors.

## 6 Conflict of Interest

*The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest*.

## 7 Author Contributions

LL and BZ: conceptualization. BZ: writing manuscript and performing analysis. PD, XS and XH: providing help in analysis. BZ, LL, XH and PD: collected the data from published literature or database. P H revised the manuscript. LL and XF supervised and supported this project.

## 8 Funding

No funding.

## 9 Acknowledgments

This study makes use of data generated by the Molecular Oncology Laboratory of Prof. Qimin Zhan, the Translational Medicine Research Center, Shanxi Medical University of Prof. Yongping Cui, the Department of Radiation Oncology, Fudan University Shanghai Cancer Center of Prof. Kuaile Zhao, the Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center of Prof. Han Liang, The Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, Prof. Norman E. Sharpless, the Institute of Clinical Pathology, Shantou University Medical College, Prof. Min Su, the Cedars-Sinai Medical Center, UCLA School of Medicine, Prof. H. Phillip Koeffler and Prof. Jie He of Cancer Institute and

Hospital Chinese Academy of Medical Sciences. We also acknowledge other Professors for sharing the fastq data, The National Center for Biotechnology Information, The European Genome-phenome Archive, The Cancer Genome Atlas for sharing the esophageal squamous cell cancer data.

The last but not the least, I want to thanks to my wife Panhong Liu, without her support, without my scientific research.

## 10 References

1. Brown J, Stepien AJ, Willem P. Landscape of copy number aberrations in esophageal squamous cell carcinoma from a high endemic region of South Africa. *BMC cancer* (2020) **20**:281. doi:10.1186/s12885-020-06788-3

2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ, He J. Cancer statistics in China, 2015. *CA: A Cancer Journal for Clinicians* (2016) **66**:115–132. doi:10.3322/caac.21338

3. Engel LS, Chow WH, Vaughan TL, Gammon MD, Risch HA, Stanford JL, Schoenberg JB, Mayne ST, Dubrow R, Rotterdam H, et al. Population attributable risks of esophageal and gastric cancers. *Journal of the National Cancer Institute* (2003) **95**:1404–1413. doi:10.1093/jnci/djg047

4. Song Y, Li L, Ou Y, Gao Z, Li E, Li X, Zhang W, Wang JJ, Xu L, Zhou Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* (2014) **508**:91–95. doi:10.1038/nature13176

5. Gao YB, Chen ZL, Li JG, Hu X da, Shi XJ, Sun ZM, Zhang F, Zhao ZR, Li ZT, Liu ZY, et al. Genetic landscape of esophageal squamous cell carcinoma. *Nature Genetics* (2014) **46**:1097–1102. doi:10.1038/ng.3076

6. Chen XX, Zhong Q, Liu Y, Yan SM, Chen ZH, Jin SZ, Xia TL, Li RY, Zhou AJ, Su Z, et al. Genomic comparison of esophageal squamous cell carcinoma and its precursor lesions by multi-region whole-exome sequencing. *Nature Communications* (2017) **8**: doi:10.1038/s41467-017-00650-0

7. Liu X, Zhang M, Ying S, Zhang C, Lin R, Zheng J, Zhang G, Tian D, Guo Y, Du C, et al. Genetic Alterations in Esophageal Tissues From Squamous Dysplasia to Carcinoma. *Gastroenterology* (2017) **153**:166–177. doi:10.1053/j.gastro.2017.03.033

8. Cui R, Kamatani Y, Takahashi A, Usami M, Hosono N, Kawaguchi T, Tsunoda T, Kamatani N, Kubo M, Nakamura Y, et al. Functional Variants in ADH1B and ALDH2 Coupled With Alcohol and Smoking Synergistically Enhance Esophageal Cancer Risk. *Gastroenterology* (2009) **137**:1768–1775. doi:10.1053/j.gastro.2009.07.070

9. Wang LD, Zhou FY, Li XMXCXM, Sun LD, Song X, Jin Y, Li JLM, Kong GQ, Qi H, Cui J, et al. Genome-wide association study of esophageal squamous cell carcinoma in chinese subjects identifies a susceptibility locus at PLCE1. *Nature Genetics* (2010) **42**:759–765. doi:10.1038/ng.648

10. Wu C, Hu Z, He Z, Jia W, Wang F, Zhou Y, Liu Z, Zhan Q, Liu Y, Yu D, et al. Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature Genetics* (2011) **43**:679–684. doi:10.1038/ng.849

11. Wu C, Kraft P, Zhai K, Chang J, Wang Z, Li Y, Hu Z, He Z, Jia W, Abnet CC, et al. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nature Genetics* (2012) **44**:1090–1097. doi:10.1038/ng.2411

12. Wu C, Wang Z, Song X, Feng XS, Abnet CC, He J, Hu N, Zuo XB, Tan W, Zhan Q, et al. Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nature Genetics* (2014) **46**:1001–1006. doi:10.1038/ng.3064

13. Lin D, Wu C, Li D, Jia W, Hu Z, Zhou Y, Yu D, Tong T, Wang M, Lin D, et al. Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nature Genetics* (2013) **45**:632–638. doi:10.1038/ng.2638

14. Chang J, Zhong R, Tian J, Li J, Zhai K, Ke J, Lou J, Chen W, Zhu B, Shen N, et al. Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. *Nature Genetics* (2018) **50**:338–343. doi:10.1038/s41588-018-0045-8

15. Hu JL, Hu XL, Lu CX, Chen XJ, Fu L, Han Q, Cang SD. Variants in the 3'-untranslated region of CUL3 is associated with risk of esophageal squamous cell carcinoma. *Journal of Cancer* (2018) **9**:3647–3650. doi:10.7150/jca.27052

16. Suo C, Yang Y, Yuan Z, Zhang T, Yang X, Qing T, Gao P, Shi L, Fan M, Cheng H, et al. Alcohol Intake Interacts with Functional Genetic Polymorphisms of Aldehyde Dehydrogenase (ALDH2) and Alcohol Dehydrogenase (ADH) to Increase Esophageal Squamous Cell Cancer Risk. *Journal of Thoracic Oncology* (2019) **14**:712–725. doi:10.1016/j.jtho.2018.12.023

17. Huang K lin, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* (2018) **173**:355-370.e14. doi:10.1016/j.cell.2018.03.039

18. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, Johann PD, Balasubramanian GP, Segura-Wang M, Brabetz S, et al. The landscape of genomic alterations across childhood cancers. *Nature* (2018) **555**:321–327. doi:10.1038/nature25480

19. Deng J, Weng X, Ye J, Zhou D, Liu Y, Zhao K. Identification of the germline mutation profile in esophageal squamous cell carcinoma by whole exome sequencing. *Frontiers in Genetics* (2019) **10**:1–10. doi:10.3389/fgene.2019.00047

20. Ko JMY, Ning L, Zhao XK, Chai AWY, Lei LC, Choi SSA, Tao L, Law S, Kwong A, Lee NPY, et al. BRCA2 loss-of-function germline mutations are associated with esophageal squamous cell carcinoma risk in Chinese. *International Journal of Cancer* (2020) **146**:1042–1051. doi:10.1002/ijc.32619

21. Knudson AG. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer* (2001) **1**:157–162. doi:10.1038/35101031

22. Kemel Y, Zhang L, Knezevic A, Patil S, Ceyhan-birsoy O, Huang K, Redzematovic A, Coskey DT, Stewart C, Pradhan N, et al. Prevalence of Germline Mutations in Cancer Susceptibility Genes in Patients With Advanced Renal Cell Carcinoma. (2018) **10065**: doi:10.1001/jamaoncol.2018.1986

23. Park S, Supek F, Lehner B. Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. *Nature Communications* (2018) **9**: doi:10.1038/s41467-018-04900-7

24. Lin DC, Hao JJ, Nagata Y, Xu L, Shang L, Meng X, Sato Y, Okuno Y, Varela AM, Ding LW, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nature Genetics* (2014) **46**:467–473. doi:10.1038/ng.2935

25. Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, Kong P, Wang JJJJ, Li YY, Chen W, Song B, et al. Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. *American Journal of Human Genetics* (2015) **96**:597–611. doi:10.1016/j.ajhg.2015.02.017

26. Hao JJ, Lin DC, Dinh HQ, Mayakonda A, Jiang YY, Chang C, Jiang Y, Lu CC, Shi ZZ, Xu X, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics* (2016) **48**:1500–1507. doi:10.1038/ng.3683

27. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, Patel N, Mlombe YB, Mulima G, Liomba NG, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI Insight* (2016) **1**:1–11. doi:10.1172/jci.insight.88755

28. Deng J, Chen H, Zhou D, Zhang J, Chen Y, Liu Q, Ai D, Zhu H, Chu L, Ren W, et al. Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. *Nature Communications* (2017) **8**:1–9. doi:10.1038/s41467-017-01730-x

29. Kim J, Bowlby R, Mungall AJ, Robertson AG, Odze RD, Cherniack AD, Shih J, Pedamallu CS, Cibulskis C, Dunford A, et al. Integrated genomic characterization of oesophageal carcinoma. *Nature* (2017) **541**:169–174. doi:10.1038/nature20805

30. Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo K v., Margulies EH. Accurate and comprehensive sequencing of personal genomes. *Genome Research* (2011) **21**:1498–1505. doi:10.1101/gr.123638.111

31. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y. SOAPnuke : A MapReduce Acceleration supported Software for integrated Quality Control and Preprocessing of High-Throughput Sequencing Data. *GigaScience* (2018) **7**:gix120. doi:10.1093/gigascience/gix120/4689118

32. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* (2010) **20**:1297–1303. doi:10.1101/gr.107524.110

33. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* (2010) **26**:589–595. doi:10.1093/bioinformatics/btp698

34. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biology* (2016) **17**:1–14. doi:10.1186/s13059-016-0974-4

35. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ. Oncotator : Cancer Variant Annotation Tool. *Human Mutation* (2015) **36**:E2423–E2429. doi:10.1002/humu.22771

36. Mayrhofer M, DiLorenzo S, Isaksson A. Patchwork: Allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biology* (2013) **14**:R24. doi:10.1186/gb-2013-14-3-r24

37. Shen R, Seshan VE. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research* (2016) **44**:1–9. doi:10.1093/nar/gkw520

38. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L, et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research* (2017) **45**:D777–D783. doi:10.1093/nar/gkw1121

39. Mirabello L, Zhu B, Koster R, Karlins E, Dean M, Yeager M, Gianferante M, Spector LG, Morton LM, Karyadi D, et al. Frequency of pathogenic germline variants in cancer-susceptibility genes in patients with osteosarcoma. *JAMA Oncology* (2020) **6**:724–734. doi:10.1001/jamaoncol.2020.0197

40. Zhao M, Sun J, Zhao Z. TSGene: A web resource for tumor suppressor genes. *Nucleic Acids Research* (2013) **41**:970–976. doi:10.1093/nar/gks937

41. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: An updated literature-based knowledgebase for Tumor Suppressor Genes. *Nucleic Acids Research* (2016) **44**:D1023–D1031. doi:10.1093/nar/gkv1268

42. Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. *Journal of Genetics and Genomics* (2017) **44**:119–121. doi:10.1016/j.jgg.2016.12.004

43. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics* (2017) **100**:267–280. doi:10.1016/j.ajhg.2017.01.004

44. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* (2015) **17**:405–423. doi:10.1038/gim.2015.30

45. Stenson PD, Mort M, Ball E v., Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The Human Gene Mutation Database: towards a comprehensive repository of

inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics* (2017) **136**:665–677. doi:10.1007/s00439-017-1779-6

46. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* (2018) **46**:D1062–D1067. doi:10.1093/nar/gkx1153

47. Lek M, Karczewski KJ, Minikel E v., Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* (2016) **536**:285–291. doi:10.1038/nature19057

48. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology* (2018) **36**:983. doi:10.1038/nbt.4235

49. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* (2009) **4**:1073–1082. doi:10.1038/nprot.2009.86

50. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nature Methods* (2010) **7**:248–249. doi:10.1038/nmeth0410-248

51. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* (2019) **47**:D886–D894. doi:10.1093/nar/gky1016

52. Geurts-Giele WRR, Leenen CHM, Dubbink HJ, Meijssen IC, Post E, Sleddens HFBM, Kuipers EJ, Goverde A, van den Ouweland AMW, van Lier MGF, et al. Somatic aberrations of mismatch repair genes as a cause of microsatellite-unstable cancers. *Journal of Pathology* (2014) **234**:548–559. doi:10.1002/path.4419

53. Cox C, Bignell G, Greenman C, Stabenau A, Warren W, Stephens P, Davies H, Watt S, Teague J, Edkins S, et al. A survey of homozygous deletions in human cancer genomes. *Proceedings of the National Academy of Sciences of the United States of America* (2005) **102**:4542–4547. doi:10.1073/pnas.0408593102

54. Ryland GL, Doyle MA, Goode D, Boyle SE, Choong DYH, Rowley SM, Li J, Bowtell DD, Tothill RW, Campbell IG, et al. Loss of heterozygosity: What is it good for? *BMC Medical Genomics* (2015) **8**:1–12. doi:10.1186/s12920-015-0123-z

55. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer. *Cancer Research* (2017) **77**:e31–e34. doi:10.1158/0008-5472.CAN-17-0337

56. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics* (2013) **92**:841–853. doi:10.1016/j.ajhg.2013.04.015

57. Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Research* (2020) doi:10.1038/s41422-020-0322-9

58. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: New online mutation analysis and recommendations to users. *Human Mutation* (2002) **19**:607–614. doi:10.1002/humu.10081

59. Dzhemileva LU, Barashkov NA, Posukh OL, Khusainova RI, Akhmetova VL, Kutuev IA, Gilyazova IR, Tadinova VN, Fedorova SA, Khidiyatova IM, et al. Carrier frequency of GJB2 gene mutations c.35delG, c.235delC and c.167delT among the populations of Eurasia. *Journal of Human Genetics* (2010) **55**:749–754. doi:10.1038/jhg.2010.101

60. Kwong A, Shin VY, Ho JCW, Kang E, Nakamura S, Teo SH, Lee ASG, Sng JH, Ginsburg OM, Kurian AW, et al. Comprehensive spectrum of BRCA1 and BRCA2 deleterious mutations in breast cancer in Asian countries. *Journal of Medical Genetics* (2016) **53**:15–23. doi:10.1136/jmedgenet-2015-103132

61. Waszak SM, Northcott PA, Buchhalter I, Robinson GW, Sutter C, Groebner S, Grund KB, Brugières L, Jones DTW, Pajtler KW, et al. Spectrum and prevalence of genetic predisposition in medulloblastoma: a retrospective genetic study and prospective validation in a clinical trial cohort. *The Lancet Oncology* (2018) **19**:785–798. doi:10.1016/S1470-2045(18)30242-0

62. Wen WX, Allen J, Lai KN, Mariapun S, Hasan SN, Ng PS, Lee DSC, Lee SY, Yoon SY, Lim J, et al. Inherited mutations in BRCA1 and BRCA2 in an unselected multiethnic cohort of Asian patients with breast cancer and healthy controls from Malaysia. *Journal of Medical Genetics* (2018) **55**:97–103. doi:10.1136/jmedgenet-2017-104947

63. Aretz S, Tricarico R, Papi L, Spier I, Pin E, Horpaopan S, Cordisco EL, Pedroni M, Stienen D, Gentile A, et al. MUTYH-associated polyposis (MAP): Evidence for the origin of the common European mutations p.Tyr179Cys and p.Gly396Asp by founder events. *European Journal of Human Genetics* (2014) **22**:923–929. doi:10.1038/ejhg.2012.309

64. Taki K, Sato Y, Nomura S, Ashihara Y, Kita M, Tajima I, Sugano K, Arai M. Mutation analysis of MUTYH in Japanese colorectal adenomatous polyposis patients. *Familial Cancer* (2016) **15**:261–265. doi:10.1007/s10689-015-9857-1

65. Klift HM van der, Tops ÃCMJ, Bik EC, Boogaard MW, Borgstein A, Hansson KBM, Ausems MGEM, Garcia EG, Green A, Hes FJ, et al. Quantification of Sequence Exchange Events between PMS2 and PMS2CL Provides a Basis for Improved Mutation Scanning of Lynch Syndrome Patients. *Human Mutation* (2010) **31**:578–587. doi:10.1002/humu.21229

66. Zhang P, Kitchen-Smith I, Xiong L, Stracquadanio G, Brown K, Richter P, Wallace M, Bond E, Sahgal N, Moore S, et al. Germline and somatic genetic variants in the p53 pathway interact to affect cancer risk, progression and drug response. *bioRxiv* (2019)835918. doi:10.1101/835918

67. Staninova-Stojovska M, Matevska-Geskovska N, Panovski M, Angelovska B, Mitrevski N, Ristevski M, Jovanovic R, Dimovski A. Molecular Basis of Inherited Colorectal Carcinomas

in the Macedonian Population: An Update. *Balkan J Med Genet* (2019) **22**:5–16. doi:10.2478/bjmg-2019-0025

68. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, et al. A global reference for human genetic variation. *Nature* (2015) **526**:68–74. doi:10.1038/nature15393

69. Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* (2010) **138**:2073-2087.e3. doi:10.1053/j.gastro.2009.12.064

70. Sourrouille I, Coulet F, Lefevre JH, Colas C, Eyries M, Svrcek M, Bardier-Dupas A, Parc Y, Soubrier F. Somatic mosaicism and double somatic hits can lead to MSI colorectal tumors. *Familial Cancer* (2013) **12**:27–33. doi:10.1007/s10689-012-9568-9

71. Ceccaldi R, Sarangi P, D'Andrea AD. The Fanconi anaemia pathway: new players and new functions. *Nature reviews Molecular cell biology* (2016) **17**:337. doi:10.1038/nrm.2016.48

72. Joshi Niraj, Anniina Färkkilä, D'Andrea AD. The Fanconi Anemia Pathway in Cancer. *Annu Rev Cancer Biol* (2019) **3**:457–478. doi:10.1016/j.physbeh.2017.03.040

73. Hsieh P, Yamane K. DNA mismatch repair: Molecular mechanism, cancer, and ageing. *Mech Ageing Dev* (2008) **129**:391–407. doi:10.1016/j.mad.2008.02.012

74. Li GM. Mechanisms and functions of DNA mismatch repair. *Cell Research* (2008) **18**:85–98. doi:10.1038/cr.2007.115

75. Li X, Heyer W-D. Homologous recombination in DNA reapir and DNA tolerance. *Cell Research* (2008) **18**:99–113. doi:10.1038/cr.2008.1

76. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MDM, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C, et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature Communications* (2014) **5**:1–14. doi:10.1038/ncomms4156

77. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* (2018) **173**:321-337.e10. doi:10.1016/j.cell.2018.03.035

78. Pavletich NP, Chambers KA, Pabo CO. The DNA-binding domain of 53 contains the four conserved regions the major mutation hot spots. *Genes & development* (1993) **7**:2556–2564.

79. Harms KL, Chen X. The functional domains in p53 family proteins exhibit both common and distinct properties. *Cell Death and Differentiation* (2006) **13**:890–897. doi:10.1038/sj.cdd.4401904

80. Baugh EH, Ke H, Levine AJ, Bonneau RA, Chan CS. Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death and Differentiation* (2018) **25**:154–160. doi:10.1038/cdd.2017.180

81. Parrales A, Iwakuma T. Targeting oncogenic mutant p53 for cancer therapy. *Frontiers in Oncology* (2015) **5**:1–13. doi:10.3389/fonc.2015.00288

82. Pradhan MR, Siau JW, Kannan S, Nguyen MN, Ouaray Z, Kwoh CK, Lane DP, Ghadessy F, Verma S. Simulations of mutant p53 DNA binding domains reveal a novel druggable pocket. *Nucleic Acids Research* (2019) **47**:1637–1652. doi:10.1093/nar/gky1314

83. Kurose K, Gilley K, Matsumoto S, Watson PH, Zhou XP, Eng C. Frequent somatic mutations in PTEN and TP53 are mutually exclusive in the stroma of breast carcinomas. *Nature Genetics* (2002) **32**:355–357. doi:10.1038/ng1013

84. Liu Y, Chen C, Xu Z, Scuoppo C, Rillahan CD, Gao J, Spitzer B, Bosbach B, Kastenhuber ER, Baslan T, et al. Deletions linked to TP53 loss drive cancer through p53-independent mechanisms. *Nature* (2016) **531**:471–475. doi:10.1038/nature17157

85. Buisson R, Dion-Côté A-M, Coulombe Y, Launay H. Cooperation of breast cancer proteins PALB2 and piccolo BRAC2 in stimulating homologous recombination. *Nature structural & molecular biology* (2010) **17**:1247–1254. doi:10.1038/nsmb.1915

86. Girardi F, Barnes DR, Barrowdale D, Frost D, Brady AF, Miller C, Henderson A, Donaldson A, Murray A, Brewer C, et al. Risks of breast or ovarian cancer in BRCA1 or BRCA2 predictive test negatives: findings from the EMBRACE study. *Genetics in Medicine* (2018) **20**:1575–1582. doi:10.1038/gim.2018.44

87. Hu N, Wang C, Han XY, He LJ, Tang ZZ, Giffen C, Emmert-Buck MR, Goldstein AM, Taylor PR. Evaluation of BRCA2 in the genetic susceptibility of familial esophageal cancer. *Oncogene* (2004) **23**:852–858. doi:10.1038/sj.onc.1207150

88. Akbari MR, Malekzadeh R, Nasrollahzadeh D, Amanian D, Islami F, Li S, Zandvakili I, Shakeri R, Sotoudeh M, Aghcheli K, et al. Germline BRCA2 mutations and the risk of esophageal squamous cell carcinoma. *Oncogene* (2008) **27**:1290–1296. doi:10.1038/sj.onc.1210739

89. Skoulidis F, Cassidy LD, Pisupati V, Jonasson JG, Bjarnason H, Eyfjord JE, Karreth FA, Lim M, Barber LM, Clatworthy SA, et al. Germline Brca2 Heterozygosity Promotes KrasG12D -Driven carcinogenesis in a murine model of familial pancreatic cancer. *Cancer Cell* (2010) **18**:499–509. doi:10.1016/j.ccr.2010.10.015

90. Arora A, Agarwal D, Abdel-Fatah TMA, Lu H, Croteau DL, Moseley P, Aleskandarany MA, Green AR, Ball G, Rakha EA, et al. RECQL4 helicase has oncogenic potential in sporadic breast cancers. *Journal of Pathology* (2016) **238**:495–501. doi:10.1002/path.4681

91. Zhao L. Mismatch repair protein expression in patients with stage II and III sporadic colorectal cancer. *Oncology Letters* (2018) **15**:8053–8061. doi:10.3892/ol.2018.8337

92. Vogelsang M, Wang Y, Veber N, Mwapagha LM, Parker MI. The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. *PLoS ONE* (2012) **7**:1–10. doi:10.1371/journal.pone.0036962

93. Haraldsdottir S, Hampel H, Tomsic J, Frankel WL, Pearlman R, de La Chapelle A, Pritchard CC. Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology* (2014) **147**:1308-1316.e1. doi:10.1053/j.gastro.2014.08.041

94. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, Lai C, Brockman D, Philippakis A, Ellinor PT, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications* (2020) **11**:1–9. doi:10.1038/s41467-020-17374-3

95. Werness BA, Parvatiyar P, Ramus SJ, Whittemore AS, Garlinghouse-Jones K, Oakley-Girvan I, DiCioccio RA, Wiest J, Tsukada Y, Ponder BAJ, et al. Ovarian carcinoma in situ with germline BRCA1 mutation and loss of heterozygosity at BRCA1 and TP53. *Journal of the National Cancer Institute* (2000) **92**:1088–1091. doi:10.1093/jnci/92.13.1088

## 11 Figure Legends

**Figure 1.** The frequency and distribution of cancer susceptibility genes (CSGs) with more than one pathogenic/likely pathogenic germline mutation detected in patients with esophageal squamous cell cancer (ESCC). Only tumor suppressor genes with more than five mutations are shown. Upper bars represent the cumulative mutation numbers of each sample. Bottom bars represent the clinical information (race, gender, age, and survival/death) about the patients. The left table presents the frequency of mutations shown in the noncancer Genome Aggregation Database (gnomAD) and the China Metabolic Analytics Project (ChinaMAP) database. Right bars represent the mutation counts. The classification of the CSG is next to the mutation name (gene name + reference SNP number or gene name + chromosome position + nucleotide change).

**Figure 2.** The distribution of pathogenic/likely pathogenic germline mutations, somatic mutations, and allele loss somatic copy number variations (SCNVs) in esophageal squamous cell cancer (ESCC) cases with potential double-hit events. Upper bars represent the clinical information (age and race) about those patients. Squares represent somatic mutations, triangles represent germline mutations, and circles represent allele loss SCNVs.

**Figure 3.** The two types of double-hit events. **(A)** The paradigm of double-hit events. **(B)** The correlation between age and double-hit event type in esophageal squamous cell cancer (ESCC) cases. The position of line is the median age, and the position of rhombus is the mean age in specific ESCC cohorts. The digits in the boxes are the numbers of ESCC cases in each category.

**Figure 4.** Significantly enriched pathways and networks in esophageal squamous cell cancer (ESCC). **(A)** The network composed of genes involved in the top 10 pathways in the Kyoto Encyclopedia of Genes and Genomes pathway enrichment. The red dots represent genes, and the blue circles represent pathways. The larger the area, the higher the degree of enrichment. The different lines represent various categories of pathways; green lines indicate genetic information processing, and purple lines indicate human disease. Solid-line rectangles and solid rectangles in blue, red, and yellow represent various gene classifications. **(B)** The x axis represents cancer susceptibility genes mutated in the Fanconi anemia pathway; the y axis represents the number of patients affected in our cohort. Red font: tumor-suppressor genes.

**Table 1.** Significance of *TP53*, *BRCA2*, and *RECQL4* pathogenic or likely pathogenic variants for ESCC risk in Chinese patients

| Gene | Chinese ESCC cohort(n=424) | | 1000 Genomes EAS(n=504) | | | | ChinaMAP (n=10,558)[2] | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $P_{burden}$ | Cases[1] (n = 424) | Controls (n = 504) | $P$[3] | OR | 95%CI | Controls (n = 10,558) | $P$ | OR | 95%CI |

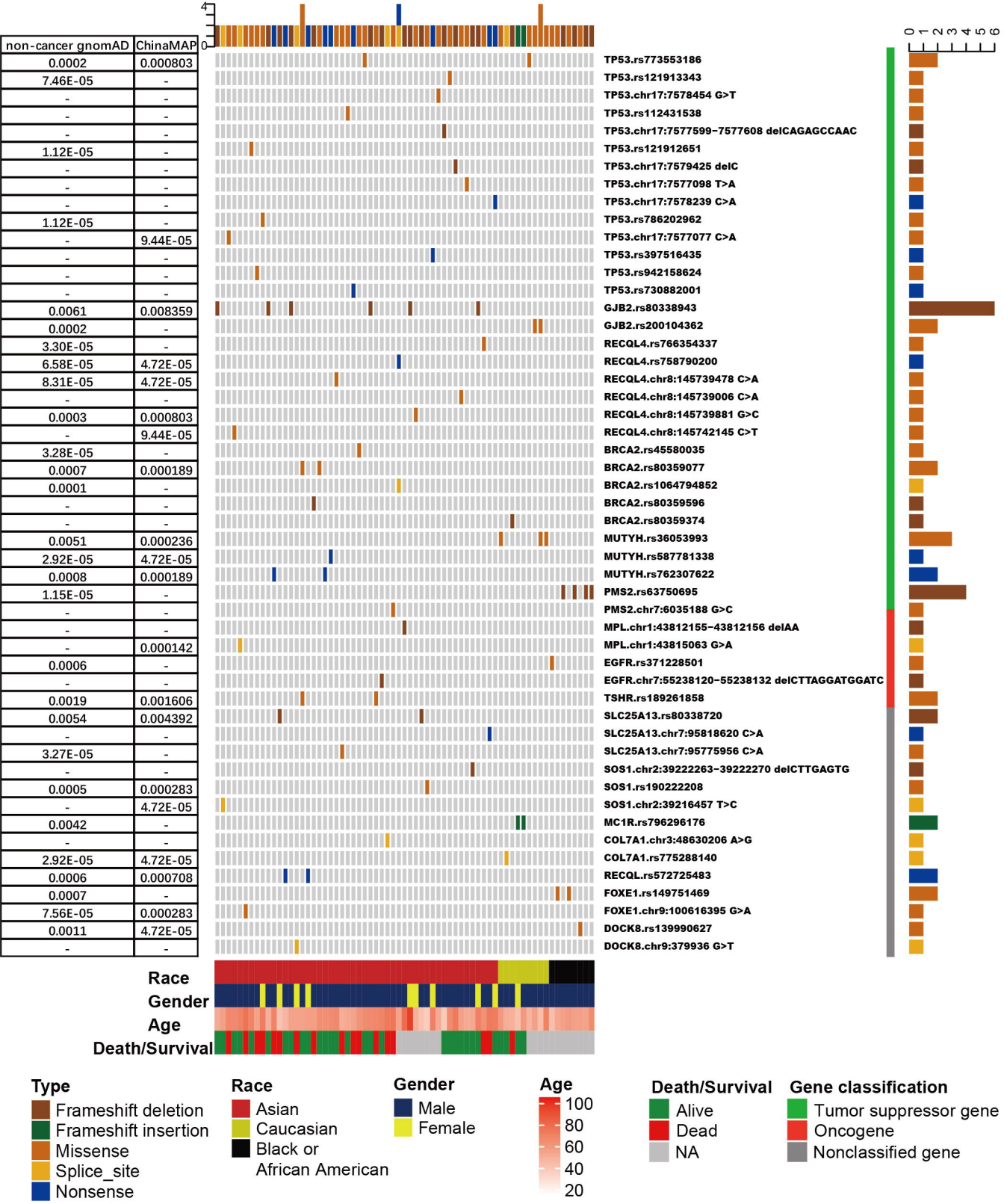
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *TP53* | $3.050 \times 10^{-3}$ | 14 (3.30%) | 4 (0.79%) | $7.359 \times 10^{-3}$ | 4.26 | 1.33 to 17.91 | 34 (0.32%) | $1.851 \times 10^{-9}$ | 10.59 | 5.21 to 20.45 |
| *BRCA2* | 0.015 | 5 (1.18%) | 0 (0%) | 0.0197 | Inf | 1.09 to Inf | 47 (0.44%) | 0.0489 | 2.68 | 0.83 to 6.75 |
| *RECQL4* | 0.035 | 6 (1.14%) | 1 (0.20%) | 0.0519 | 7.21 | 0.87 to 332.23 | 41 (0.39%) | 0.0089 | 3.69 | 1.27 to 8.81 |

Abbreviations: ChinaMAP, China Metabolic Analytics Project; EAS, East Asian; ESCC, esophageal squamous cell cancer; OR, odd ratio; CI, confidence interval; Inf, infinity.

[1]Mutation annotation are based on *TP53* transcript: NM_001126112, *BRCA2* transcript: NM_000059 and *RECQL4* transcript: NM_004260.

[2]ChinaMAP, *TP53*, *BRCA2*, and *RECQL4* variants were exported from http://www.mbiobank.com/ on June 2, 2020.

[3]Fisher exact test.


non-cancer gnomAD
ChinaMAP
0.0002 0.000803
7.46E-05 -
- -
- -
- -
1.12E-05 -
- -
- -
- -
1.12E-05 -
- 9.44E-05
- -
- -
- -
0.0061 0.008359
0.0002 -
3.30E-05 -
6.58E-05 4.72E-05
8.31E-05 4.72E-05
- -
0.0003 0.000803
- 9.44E-05
3.28E-05 -
0.0007 0.000189
0.0001 -
- -
- -
0.0051 0.000236
2.92E-05 4.72E-05
0.0008 0.000189
1.15E-05 -
- -
- -
- 0.000142
0.0006 -
- -
0.0019 0.001606
0.0054 0.004392
- -
3.27E-05 -
- -
0.0005 0.000283
- 4.72E-05
0.0042 -
- -
2.92E-05 4.72E-05
0.0006 0.000708
0.0007 -
7.56E-05 0.000283
0.0011 4.72E-05
- -
TP53.rs773553186
TP53.rs121913343
TP53.chr17:7578454 G>T
TP53.rs112431538
TP53.chr17:7577599-7577608 delCAGAGCCAAC
TP53.rs121912651
TP53.chr17:7579425 delC
TP53.chr17:7577098 T>A
TP53.chr17:7578239 C>A
TP53.rs786202962
TP53.chr17:7577077 C>A
TP53.rs397516435
TP53.rs942158624
TP53.rs730882001
GJB2.rs80338943
GJB2.rs200104362
RECQL4.rs766354337
RECQL4.rs758790200
RECQL4.chr8:145739478 C>A
RECQL4.chr8:145739006 C>A
RECQL4.chr8:145739881 G>C
RECQL4.chr8:145742145 C>T
BRCA2.rs45580035
BRCA2.rs80359077
BRCA2.rs1064794852
BRCA2.rs80359596
BRCA2.rs80359374
MUTYH.rs36053993
MUTYH.rs587781338
MUTYH.rs762307622
PMS2.rs63750695
PMS2.chr7:6035188 G>C
MPL.chr1:43812155-43812156 delAA
MPL.chr1:43815063 G>A
EGFR.rs371228501
EGFR.chr7:55238120-55238132 delCTTAGGATGGATC
TSHR.rs189261858
SLC25A13.rs80338720
SLC25A13.chr7:95818620 C>A
SLC25A13.chr7:95775956 C>A
SOS1.chr2:39222263-39222270 delCTTGAGTG
SOS1.rs190222208
SOS1.chr2:39216457 T>C
MC1R.rs796296176
COL7A1.chr3:48630206 A>G
COL7A1.rs775288140
RECQL.rs572725483
FOXE1.rs149751469
FOXE1.chr9:100616395 G>A
DOCK8.rs139990627
DOCK8.chr9:379936 G>T
Race
Gender
Age
Death/Survival
Type
Frameshift deletion
Frameshift insertion
Missense
Splice_site
Nonsense
Race
Asian
Caucasian
Black or African American
Gender
Male
Female
Age
100
80
60
40
20
Death/Survival
Alive
Dead
NA
Gene classification
Tumor suppressor gene
Oncogene
Nonclassified gene

Age
Race
TP53
GJB2
FAT1
CDKN2A
BRCA2
BRCA1
KMT2D
RB1
MUTYH
FANCG
ATM
RAD51D
CHEK2
SAMD9L
ERBB4
FANCD2
PTEN
POT1

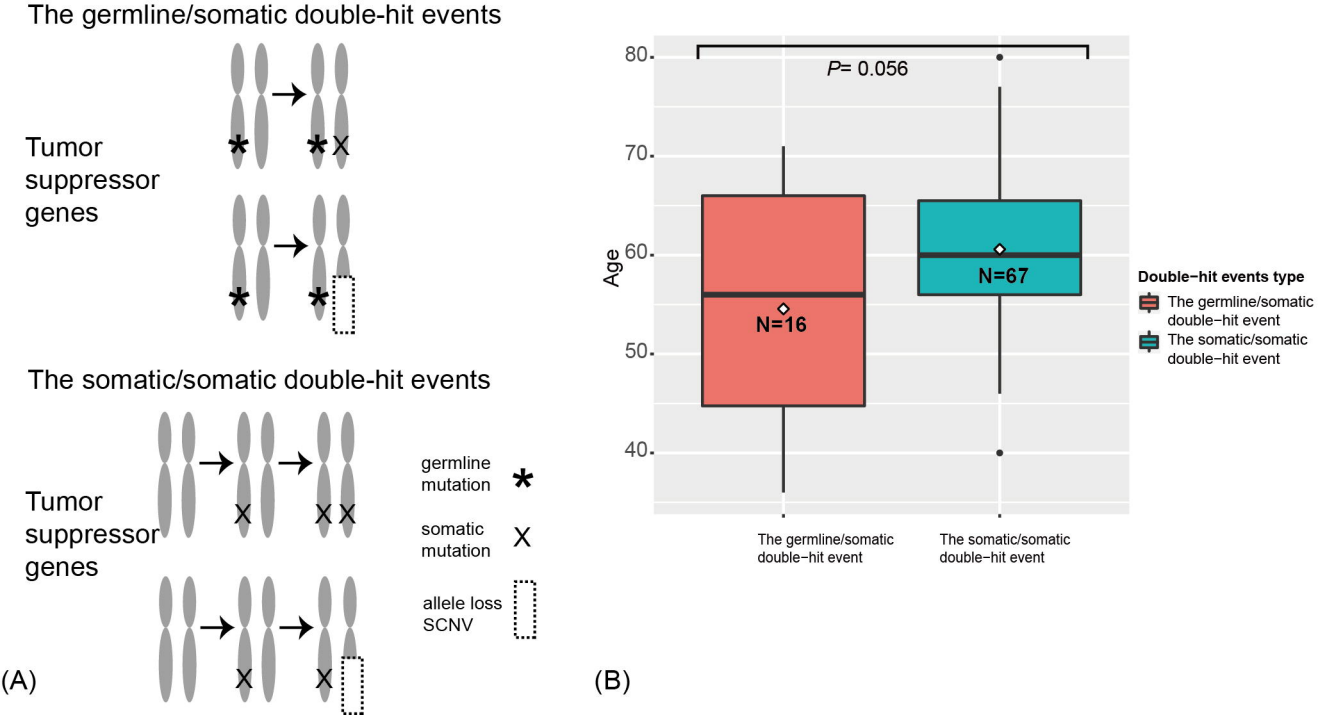


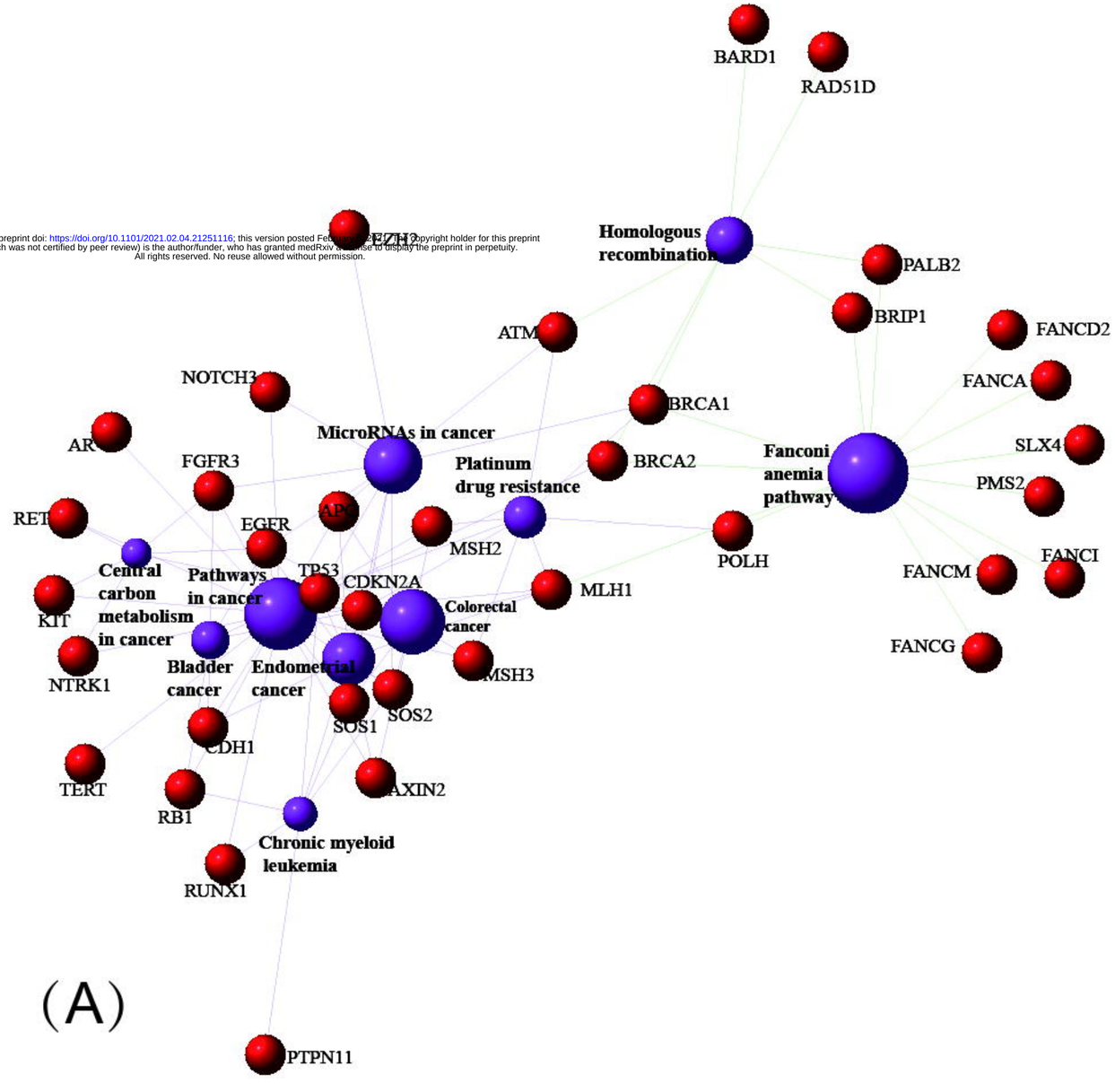
Mutation
- Allele loss SCNV
- Nonsense
- Splice_site
- Frameshift deletion
- Frameshift insertion
- Missense
- Double mutations

Type
- Somatic mutation
- Germline mutation
- Allele loss SCNV

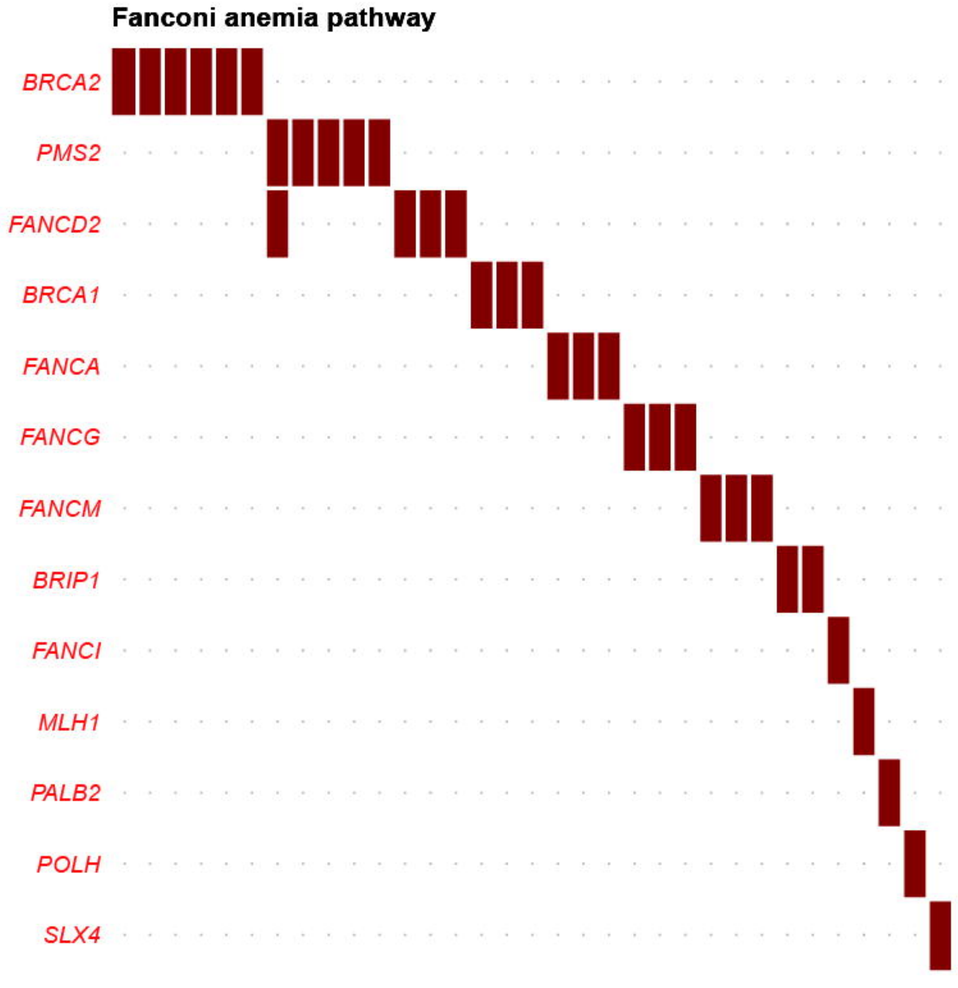
Age
80
70
60
50
40

Race
- Black or African American
- Asian
- Caucasian


The germline/somatic double-hit events
Tumor suppressor genes
The somatic/somatic double-hit events
Tumor suppressor genes
germline mutation
somatic mutation
allele loss SCNV
(A)
P= 0.056
Age
80
70
60
50
40
N=16
N=67
The germline/somatic double−hit event
The somatic/somatic double−hit event
Double−hit events type
The germline/somatic double−hit event
The somatic/somatic double−hit event
(B)




(A)




(B)