Published in final edited form as:
Int J Cancer. 2008 October 1; 123(7): 1610–1615. doi:10.1002/ijc.23682.

Replication of a genome-wide case-control study of esophageal squamous cell carcinoma

David Ng[1], **Nan Hu**[1], **Ying Hu**[2], **Chaoyu Wang**[1], **Carol Giffen**[3], **Ze-Zhong Tang**[4], **Xiao-You Han**[4], **Howard H. Yang**[2], **Maxwell P. Lee**[2], **Alisa M. Goldstein**[1], and **Philip R. Taylor**[1]

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, Maryland [2] Laboratory of Population Genetics, Center for Cancer Research, National Cancer Institute, NIH, DHHS, Bethesda, Maryland [3] Information Management Service, Inc., Silver Spring, Maryland [4] Shanxi Cancer Hospital, Taiyuan, Shanxi, People's Republic of China

Abstract

In a previous pilot case-control study of individuals diagnosed with esophageal squamous cell carcinoma (ESCC) and matched controls from a high-risk area in China, we identified 38 single nucleotide polymorphisms (SNPs) associated with ESCC located in or near one of 33 genes. In the present study, we attempted to replicate the results of these 38 gene-related SNPs in a new sample of 300 ESCC cases and 300 matched controls from the same study conducted in Shanxi Province, China. Among 36 evaluable SNPs, four were significant in one or more analyses, including SNPs located in *EPHB1*, *PGLYRP2*, *PIK3C3*, and *SLC9A9*, although the odds ratios (ORs) for these genotypes were modest. Associations were found with *EPHB1*/rs1515366 (OR 0.92, 95% CI 0.86-0.99; $p = 0.019$), *PIK3C3*/rs52911 (OR 0.93, 95% CI 0.88-0.99; $p = 0.02$), and *PGLYRP2*/rs959117 (OR 0.93, 95% CI, 0.86-1.01; $p = 0.061$) in general linear models (additive mode); and the genotype distribution differed between cases and controls for *SLC9A9*/rs956062 ($p = 0.024$). To examine these four genes in more detail, 40 HapMap-based tag SNPs from these four genes were evaluated in the same subjects, and seven additional SNPs associated with ESCC were identified. Further confirmation of these findings in other populations and other studies are needed to determine if the signals from these SNPs are indirectly associated due to linkage disequilibrium, or are directly related to biologic function and the development of ESCC.

Keywords

Esophageal cancer; Replication study; SNP

Introduction

Esophageal squamous cell carcinoma is one of the most common malignancies in China where there is also substantial geographic variation in its incidence. Shanxi Province in north central China has one of the highest rates of ESCC in China and the world. The standardized mortality rate for ESCC in Shanxi province is greater than 100/100,000 person-years.[1], [2] The etiology of this high incidence of ESCC is only incompletely understood. Past

Correspondence: David Ng, M.D. Division of Cancer Epidemiology and Genetics National Cancer Institute 6120 Executive Boulevard, Rm 7112 Bethesda, MD 20892-7236 Tel. 301-594-7931 Fax: 301-402-4489 email: davidng@mail.nih.gov.

Statement: This paper describes our findings from a replication study of 38 gene-related SNPs associated with esophageal squamous cell carcinoma (ESCC) identified from a pilot whole genome-wide association study. Four out of 38 SNPs replicated their association with ESCC; further SNP genotyping of the four candidate genes identified seven additional ESCC associated SNPs.

studies suggested that family history of ESCC,[3,4] low levels of selenium[5] and vitamin E,[6] as well as tooth loss[7] were associated with a higher risk of ESCC. Within high-risk regions of China, there is also strong evidence of familial aggregation[3,4] suggesting an inherited component to ESCC risk.

Past studies suggest evidence of a major gene contributing to ESCC risk in high-incidence regions of China. Segregation analysis of 221 ESCC Chinese families from Linxian supported the presence of an autosomal recessive gene with an estimated frequency of 19% contributing 4% of total ESCC cases.[8] In the Shanxi population, some ESCC families exhibited an apparent dominant mode of inheritance, although most do not appear to follow a mendelian pattern. Therefore, familial aggregation maybe due to both shared environmental and genetic factor(s). ESCC is most likely a complex disease with multiple genetic loci, with each loci contributing a small percentage to the overall risk. One challenge in studying a complex disease is insufficient power to detect low-risk susceptibility genes with rare variants. Studying populations in high-risk regions, such as north central China, is one strategy to identify such low level susceptibility genes.

As an initial step in the identification of ESCC susceptibility loci, we performed a whole-genome scan using the Affymetrix 10K SNP array. This pilot case-control study included 50 male ESCC cases and 50 matched controls from Shanxi province.[9] After adjustment for covariates (family history, alcohol use, tobacco use, pickled vegetable consumption, and age) in a general linear model (GLM), 38 significant gene-related single nucleotide polymorphisms (SNPs) were identified.[9] The primary aim of the current study, therefore, was to determine whether any of the 38 gene-related SNPs associated with ESCC in the pilot study replicated in a new and larger set of cases and controls. In phase 1 of the study, we genotyped these same 38 gene-related SNPs in 300 new male ESCC cases and their matched controls. In phase 2, HapMap-based tag SNPs from the four genes with significantly-associated SNPs in the first phase were reexamined in the same cases and controls.

Materials and Methods

Study population

Newly diagnosed ESCC cases and matched controls were identified and enrolled for a study of esophageal cancer between 1998 and 2004 at the Shanxi Cancer Hospital in Taiyuan, Shanxi Province, People's Republic of China. Inclusion criteria for this study were; 1) residence in one of five selected geographic areas surrounding and including Taiyuan; 2) no prior history of treatment for ESCC; 3) election of surgical treatment for ESCC; and 4) willingness to participate in the study. Individuals who did not meet all four criteria were excluded from the study. Age, sex, and neighborhood-matched controls were selected and interviewed within six months of each case diagnosed with ESCC. Shanxi province was the ancestral home for all participants. Following enrollment, patients and controls had blood samples drawn and were interviewed to obtain demographic and lifestyle histories related to cancer risk (i.e., age, tobacco and alcohol use, pickled vegetable consumption, and family history of upper gastrointestinal [UGI] tract cancer). Family history questionnaire obtained information on 1st, 2nd, 3rd degree blood relatives and spouses with a history of UGI cancer. From among this group, 300 males diagnosed with ESCC and 300 matched controls were selected for this replication SNP association study. This study was approved by the institutional review boards of the Shanxi Cancer Hospital and the US National Cancer Institute and written consent was obtained from all participants.

Genotyping

Genomic DNA was extracted and purified from venous blood drawn from cases and controls by standard methods. In the first phase of the current study, multiplex oligonucleotide ligation assays (MOLA) were used to genotype the 38 significant gene-related SNPs identified in the pilot study. MOLA is a multi-step protocol which uses a modification of the ligase-mediated gene detection assay described by Landegren *et al*.10 For the initial step, we queried the NCBI dbSNP data base to define the genomic regions which contained the 38 gene-related SNPs associated with ESCC in the pilot study.9 Primers were designed to perform multiplex (up to 14 reactions from a single sample) polymerase chain reaction (PCR) amplification of these regions using Taq Gold (Applied Biosystems, Foster City, CA) and an MJR thermal cycler (MJ Research, Waltham, MA). PCR primer sequence and cycling parameters are available upon request. PCR products were incubated with Exonuclease I and Antarctic phosphatase (New England Biolabs, Ipswich, MA) to remove the overhanging 5′ nucleotides and 5′ phosphates so that the PCR amplicons did not join together in the subsequent ligation step. Each SNP was detected using three oligonucleotide primers to form fluorescent fragments that differed in length by two base pairs to represent the alternate alleles of each SNP. The common primer was complementary to the genomic sequence 5′ upstream of the single nucleotide polymorphism and varied from 24 to 36 nucleotides in length depending on primer design. Common primers were fluorescently labeled (-fam, vic, pet, ned; Applied Biosystems, Foster City, CA) at the 5′ end and terminated at the nucleotide adjacent to and 5′ upstream of the single nucleotide polymorphism. Allele-specific primers started at the polymorphic nucleotide position of the 5′ end and terminated 26 to 34 nucleotides 3′ to this site. Prior to ligation, the allele-specific primers were incubated (MJR thermal cycler, 37°C 1 hr then inactivated at 70°C for 20 min) with T4 polynucleotide kinase (Invitrogen, Carlsbad, CA) to transfer a phosphate to the 5′-OH group of the first nucleotide of each allele-specific primer. Common primers were joined to their respective allele-specific primers in the ligation step to form a fluorescent fragment. The fluorescent fragments were separated by electrophoresis and detected by an ABI 3130XL sequencer (Applied Biosystems, Foster City, CA). Genotype calls were made using GeneMapper v4.0 software (Applied Biosystems, Foster City, CA) and call rates ranged from 98-100%.

In the second phase of the present study, SNPs associated with ESCC ($p < 0.05$) in the first phase were investigated further by selecting additional SNPs that were not in linkage disequilibrium (LD) to span the respective candidate genes. Forty-seven new gene-related SNPs were selected for MOLA genotyping in four genes (Ephrin receptor EPHB1 [*EPHB1*], n= 17; Peptidoglycan recognition protein 2 [*PGLYRP2*], n=3; Phosphatidylinositol 3-kinase, class 3 [*PIK3C3*], n=8; Solute carrier family 9, isoform A9 [*SLC9A9*], n=19) to search for additional confirmatory signals across these genes. To maximize information content and minimize genotyping cost and time, tag SNPs were selected for these genes from the International HapMap Project (http://www.hapmap.org/). Han Chinese in Beijing (CHB) genotype data for the four candidate genes was downloaded from HapMap and loaded onto Haploview (Broad Institute of MIT and Harvard, Cambridge, MA). Tagger selection algorithm (de Bakker *et al.*)11 in Haploview was used to select tag SNPs (pairwise tagging function, r^2 threshold = 0.8). In total, from phase 1 and phase 2, 51 SNPs were genotyped (*EPHB1*, n= 18; *PGLYRP2*, n=4; *PIK3C3*, n=9; *SLC9A9*, n=20) across these four genes in the 300 ESCC cases and 300 matched controls studied.

Statistical analyses

Allele frequencies for each genotype were derived from the observed distribution among controls and used to calculate expected genotype percentages. As a check on genotyping quality for each SNP, observed genotype frequencies for controls were compared to expected frequencies to test for Hardy-Weinburg equilibrium (HWE) using a chi-square test with one degree of freedom. HWE calculations were made with the online tool http://www.genes.org.uk/software/hardy-weinberg.shtml. In addition, 24 samples were sequenced for eight different SNPs to compare the genotype call rate accuracy. SNPs that did not meet HWE ($p < 0.01$) or did not have 100% genotype concordance between the MOLA assay and sequencing result were dropped from further analysis (for phase 1, n=2). Thirty-six SNPs were left for statistical analysis in the first phase of the study (Fig. 1). For quality control in phase 2, SNPs that had a call rate of < 92% and or controls not in HWE were dropped from the analysis. Seven SNPs did not pass quality control, leaving 40 evaluable SNPs.

For each SNP, distributions of genotype frequencies between cases and controls were compared using generalized linear models (GLM). These tests were run to determine the association between each SNP and ESCC. Since the inheritance mode for these genes in relation to ESCC is unknown, additive models were used for our primary GLM analyses, with genotypes coded generically as (AA, AB, BB) = (0, 1, 2). The additive model assumes a stepwise increase in risk from homozygous wildtype allele to heterozygous and homozygous with risk allele. We also performed frequency distribution analyses using 2X3 contingency table (Fisher's exact test, three genotypes each for cases and controls).

The homozygous genotype for the common allele of each SNP in the control group was used as the referent in calculating odds ratios (OR) and 95% confidence intervals (95%CI).

Significant SNPs from the GLM additive model were validated with permutation testing based on reassigning case/control status 10,000 times. As the permuted *p* values did not differ from the observed *p* values derived from the GLM analysis and the 2×3 frequency distribution tests, only the permuted *p* values are presented.

We evaluated the single most significant protective SNP from each of these four genes in multivariate GLM models where all four SNPs were evaluated simultaneously, with and without cross product terms, to assess both main effects and potential interactions.

Using the single most protective SNP from each of our four targeted genes, we also evaluated the combined effects of multiple protective alleles. Multiple protective alleles were examined in three ways: (i) as a continuous variable; (ii) categorized into four groups approximating quartiles as: group 1 (0-1 protective allele), group 2 (2 protective alleles), group 3 (3 protective alleles), and group 4 (4-6 protective alleles; there were no individuals with 7-8 protective alleles) and analyzed using indicator variables; and (iii) combined into a single ordinal variable to test for trend.

SNPs found to have a statistically significant association and located in the same gene were tested for evidence of pairwise LD by calculating D' and r^2 using Partek Genomics Suite v6.3 (Partek Inc., St. Louis, MO).

All *p* values reported are two-sided. GLM analyses and Chi-square tests used in the genotype comparisons were performed with the R program language. The same statistical analysis was applied to the 40 evaluable tag SNP genotype data from the second phase of the study.

Results

Clinical characteristics of the 600 case-control subjects are shown in Table I. Median age was 59 and tobacco and alcohol use and pickled vegetable consumption were similarly distributed between cases and controls. The only difference was a higher frequency of positive family history of UGI cancer among cases (28%) as compared to controls (18%). Of the five variables, only a positive family history of UGI cancer was significantly associated with risk (OR 1.84, 95% CI 1.25-2.72; $p = 0.002$).

Among the 38 gene-related SNPs identified in the pilot study and tested in phase 1 of this replication study, 36 SNPs passed the HWE quality control check. Of these 36 SNPs, four (*EPHB1*/rs1515366, *PGLYRP2*/rs959117, *PIK3C3*/ rs52911, and *SLC9A9*/rs956062) were statistically significantly associated with ESCC in one or more of the tests we applied: *EPHB1*/rs1515366 ($p = 0.049$), *PIK3C3*/rs52911 ($p = 0.015$), and *SLC9A9*/rs956062 ($p = 0.024$) were significant based on case-control differences in genotype frequencies (Table II); and *PIK3C3*/rs52911 (OR 0.93, 95% CI 0.88-0.99; $p = 0.02$) and *EPHB1*/rs1515366 (OR 0.92, 95% CI 0.86-0.99; $p = 0.019$) were significant in GLM additive models. ORs were decreased for the minor homozygous variant genotypes for *PIK3C3*/rs52911 (C/C, OR 0.86) and *EPHB1*/rs1515366 (A/A, OR 0.84) compared to their respective heterozygous genotypes. *PGLYRP2*/rs959117 was marginally non-significant in the GLM additive model (OR 0.93, 95% CI 0.86-1.01; $p = 0.061$). *SLC9A9*/rs956062 was not significant in the GLM analysis ($p = 0.222$).

In the second phase of this study, an additional 47 tag SNPs were selected and genotyped (Fig. 1) from the CHB HapMap data base to capture information across these four genes (*EPHB1*, *PGLYRP2*, *PIK3C3*, and *SLC9A9*). In total from phase 1 and 2, 51 SNPs were genotyped among these four genes.

Among the 17 SNPs genotyped in *EPHB1*, 15 were evaluable. Two SNPs (rs11716850, rs17712258) showed a significant difference in genotype distribution between cases and controls, however, neither SNP was significant in the GLM analysis (Table II). rs11716850 and rs17712258 are noncoding SNPs located in intron 13 and the 3′ region of *EPHB1*. Both SNPs are in strong LD ($D' = 0.99$ and $r^2 = 0.99$), however, neither SNP is in LD with rs1515366 ($D' = 0.007$, 0.009, and $r^2 < 0.001$ for both). Thus, rs11716850 and rs17712258 together, represent a second signal in *EPHB1* associated with ESCC. In total, phase 1 and 2 results identified two signals in *EPHB1* associated with ESCC.

Three additional SNPs were genotyped in *PGLYRP2* and two were evaluable. rs892145 (OR 1.06, 95% CI 1.00-1.12; $p = 0.049$) and rs4264508 (OR 0.94, 95% CI 0.89-0.99; $p = 0.017$) were both significant in GLM analysis; rs4264508 also showed a different genotype distribution ($p =0.033$) (Table II). rs892145 is a nonsynonymous coding SNP (lysine to methionine at position 270) located in exon 2; rs4264508 is in intron 2 of *PGLYRP2*. These two significant SNPs in *PGLYRP2* are not in LD ($D' =0.64$, $r^2=0.44$) with each other and represent two distinct signals associated with ESCC.

Eight additional SNPs were genotyped in *PIK3C3*, but none of the five evaluable SNPs were significantly associated with ESCC in either GLM analysis or by frequency distribution (Supplemental Table I).

Among the 19 SNPs genotyped in *SLC9A9*, 18 were evaluable, including three (rs838598, rs16853475, and rs10804689) with statistically significant differences in the distribution of their genotype frequencies; rs16853475 (OR 0.91, 95% CI 0.85-0.98; $p = 0.012$), and rs10804689 (OR 1.08, 95% CI 1.01-1.15; $p = 0.016$) were also significantly different in GLM analyses (Table II). Altogether, four SNPs in *SLC9A9* were associated with ESCC and none of these SNPs were in LD with each other.

Simultaneous inclusion of the four most significant protective SNPs (*EPHB1*/rs1515366, *PGLYRP2*/rs4264508, *PIK3C3*/rs52911, and *SLC9A9*/rs16853475) as main effects in GLM analyses indicated that each SNP was an independent predictor of ESCC risk; no evidence for effect modification on a multiplicative scale between these four SNPs was seen (data not shown). Analysis of multiple protective alleles was strongly positive: risk decreased 8% for each additional protective allele (OR 0.92, 95% CI 0.89-0.95; $p = 1.32E^{-06}$) and the trend across protective allele categories was highly significant ($p = 2.86E^{-06}$). Men with four or more protective alleles had a 26% reduced risk compared to men with zero or one (OR 0.74, 95% CI 0.66-0.84; $p = 1.53E^{-06}$; Table III).

For completeness, the GLM analyses for dominant and recessive inheritance modes for all SNPs evaluated here are shown in Supplemental Table I.

Discussion

Shanxi Province in north central China has one of the highest esophageal cancer mortality rates in the world.4 In 1979, a Chinese national cancer mortality survey conducted in Yangcheng County, Shanxi Province, found that all ESCC cases occurred in 8% of households. ESCC familial aggregation4 is striking and suggests that a subset of this population carried genetic variants resulting in ESCC susceptibility. To identify genetic factors leading to ESCC susceptibility, we conducted a replication case-control study of 38 gene-related SNPs associated with ESCC in a previous pilot study.9 Among 36 evaluable SNPs previously identified, four (11%) replicated their association with ESCC risk in the first phase of this study. Further genotyping of tag SNPs identified seven additional SNPs associated with ESCC. Altogether, four genes – *EPHB1*, *PGLYRP2*, *PIK3C3*, and *SLC9A9* – were found that contained one or more SNPs associated with ESCC. Six out of 11 SNPs identified in this study were significant in the GLM analysis (Table II). Choosing the common allele as the referent, four SNPs (rs1515366, rs4264508, rs52911, rs16853475) from the GLM revealed the minor allele as protective and the major allele as the risk allele for ESCC susceptibility. The exceptions were rs892145 and rs10804689, where the variant was the risk allele for ESCC. It is unlikely that each allele from these six SNPs by themselves is of sufficient magnitude to affect risk at the population level, but protection (or risk) associated with having a combination of protective (risk) alleles is substantially greater, and warrants further consideration in determining genetic susceptibility at the population level.

These four genes have interesting and diverse functions. *EPHB1* has several functions related to cell growth, migration and angiogenesis, all key regulatory points critical in the multi-step evolution of tumorigenesis and metastasis. *PGLYRP2* is part of the innate immune response. *PIK3C3* is a member of the phosphoinositide 3-kinase (PI3K) family of proteins that regulate cell survival, migration and growth. *SLC9A9* mediates sodium/ hydrogen exchange and regulates intracellular pH.

EPHB1 is a member of the Ephrin family of receptors. Ephrin receptors are transmembrane tyrosine kinases that activate signal transduction pathways to regulate and guide cell migration, and mediate cell repulsion to prevent invasion of cells.12 The ephrin signaling pathway interacts with the WNT pathway to regulate cell growth during embryogenesis, tissue regeneration and carcinogenesis.13 *EPHB1* has been studied in colorectal cancer where prior studies showed that a loss of ephrinB receptor activity in mice accelerated the formation of adenocarcinomas.14 Furthermore, *EPHB1* appears to have a role in angiogenesis and oncogenesis mediated through its actions in cell adhesion and chemoattraction.15 The association noted here between *EPHB1* and ESCC is the first reported in the literature and makes it an interesting candidate gene; however, its role in ESCC is yet to be determined.

PGLYRP2, is a member of the peptidoglycan recognition family of proteins that are highly conserved between vertebrates and invertebrates. These proteins are part of the innate immune system, an early defense mechanism that recognizes bacteria through their cell wall component, peptidoglycan.16 In humans, *PGLYRP2*, is constitutively expressed in the liver16 and secreted into the circulation.17 *PGLYRP2* is expressed at lower levels16 in other parts of the GI tract (i.e., colon, stomach, pancreas). However, its expression is inducible in other tissues of the human body upon exposure to bacteria.18 *PGLYRP2* and ESCC have not been studied, however, its expression has been shown to be down-regulated in colorectal cancer,19 and hepatoblastoma20 compared to normal colonic and hepatic tissues, respectively. In light of the causal relation between early *Helicobacter pylori* infection and gastric cancer, an interesting hypothesis generated by this *PGLYRP2* finding is whether there is an infectious component to ESCC susceptibility in which innate immune response modulates individual risk.

PIK3C3 is a member of the PI3K family. Members of this protein family are involved in receptor-mediated signal transduction and intracellular vesicle trafficking.21 *PIK3C3* generates phosphatidylinositol-3-phosphate which activates *AKT1*.22 PI3K/AKT pathway is a regulator of cell proliferation and apoptosis.23 An expression study of t(12;21)(p13;q22) chromosomal translocation in B-cell acute lymphocytic leukemia found a correlation with up-regulation of *PIK3C3* and increased tumor cell survival.24

SLC9A9 is a member of a family of integral membrane proteins that mediate sodium/hydrogen exchange (NHE) to maintain intracellular pH, electrolyte, and volume homeostasis.25 NHE proteins regulate many cellular functions such as cell motility and cardiac myocyte contraction. Malignant cells of diverse origin have elevated intracellular pH.26 Elevation of intracellular pH is an integral feature of cellular transformation to malignancy and is mediated by *SLC9A1* (NHE1 [sodium/hydrogen exchanger isoform 1]).26 Although the role of *SLC9A9* in carcinogenesis has not been elucidated, recent studies showed that *SLC9A9* is down-regulated in hormone-refractory prostate cancer compared to hormone-sensitive prostate cancer.27

This is the first SNP replication study of ESCC susceptibility to follow from a prior whole-genome wide association study. Strengths of this study include that it was conducted in a high-risk, highly homogeneous population and that controls were well-matched and had comparable environmental exposures. Furthermore, this is a replication study; therefore, findings should be more robust than results from an initial discovery study. We applied stringent genotyping quality control as a check on genotyping accuracy. For associated SNPs, the permuted *p* values for the GLM analysis and 2×3 distributions were not statistically different from observed *p* values, lessening the probability that these are false positive findings. A shortcoming of the study is that not all SNPs in the four candidate genes evaluated were genotyped. For genes that span a large distance and exhibit low LD such as *SLC9A9* and *EPHB1*, it simply was not feasible to genotype the 250+ tag SNPs required to cover all LD blocks. Thus, inevitably we may have missed associated SNPs. Although 300 cases is the largest number of subjects in an ESCC SNP association replication study to date, it is still a relatively small study and multiple comparisons were made. Finally, this study was conducted only in Chinese males and generalizability to females and other populations remain to be determined.

Association studies do not permit a determination of whether findings are indirect (i.e., secondary to LD with a region of true biologic importance) or direct biomarkers of biologic relevance for the disease under study. Additional replications of the findings from these studies are needed to further validate the results in this population, as well as to extend conclusions to women and other high-risk populations elsewhere in the world. Investigation of the functional relevance of these four candidate genes are also needed, and might include comparisons of gene expression levels by genotypes for those SNPs that differ between cases and controls, and/or sequencing upstream promoter regions to identify functional SNPs.

In conclusion, this replication study confirmed previously-identified associations for SNPs in four genes with ESCC risk. Additional studies to further replicate and generalize these findings as well as determine functionality are needed to define the role of these genes in the development of ESCC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics and Center for Cancer Research. We would like to thank the generosity of the Shanxi Cancer Hospital participants without whom this study would not have been possible.

Abbreviations

95% CI	95% confidence interval
CHB	Han Chinese in Beijing
dbSNP	single nucleotide polymorphism database
D'	D-prime
E^{-06}	10 to the −6 power
EPHB1	Ephrin receptor EPHB1
ESCC	esophageal squamous cell carcinoma
GI	gastrointestinal
GLM	general linear model
HapMap	International HapMap Project
HWE	Hardy-Weinburg equilibrium
LD	linkage disequilibrium
MOLA	multiplex oligonucleotide ligation assay
NCBI	National Center for Biotechnology Information
OR	odds ratio
P	P-value
PCR	polymerase chain reaction, PGLYRP2, peptidoglycan recognition protein 2
PIK3C3	phosphatidylinositol 3-kinase, class 3
r^2	r-squared
SLC9A9	solute carrier family 9, isoform A9
SNPs	single nucleotide polymorphisms
WNT	wingless-type gene family

References

1. Li JY. Epidemiology of esophageal cancer in China. Natl Cancer Inst Monogr. 1982; 62:113–20. [PubMed: 7167171]
2. National Cancer Control Office. Investigation of cancer mortality in China. People's Health Publishing House; Beijing: 1980.
3. Hu N, Dawsey SM, Wu M, Taylor PR. Family history of oesophageal cancer in Shanxi Province, China. Eur J Cancer. 1991; 27:1336. [PubMed: 1835609]
4. Hu N, Dawsey SM, Wu M, Bonney GE, He LJ, Han XY, Fu M, Taylor PR. Familial aggregation of oesophageal cancer in Yangcheng County, Shanxi Province, China. Int J Epidemiol. 1992; 21:877–82. [PubMed: 1468848]
5. Mark SD, Qiao YL, Dawsey SM, Wu YP, Katki H, Gunter EW, Fraumeni JF Jr, Blot WJ, Dong ZW, Taylor PR. Prospective study of serum selenium levels and incident esophageal and gastric cancers. J Natl Cancer Inst. 2000; 92:1753–63. [PubMed: 11058618]
6. Taylor PR, Qiao YL, Abnet CC, Dawsey SM, Yang CS, Gunter EW, Wang W, Bot WJ, Dong ZW, Mark SD. Prospective study of serum vitamin E levels and esophageal and gastric cancers. J Natl Cancer Inst. 2003; 95:1414–6. [PubMed: 13130117]
7. Abnet CC, Qiao YL, Mark SD, Dong ZW, Taylor PR, Dawsey SM. Prospective study of tooth loss and incident esophageal and gastric cancers in China. Cancer Causes Control. 2001; 12:847–54. [PubMed: 11714113]
8. Carter CL, Hu N, Wu M, Lin PZ, Murigande C, Bonney GE. Segregation analysis of esophageal cancer in 221 high-risk Chinese families. J Natl Cancer Inst. 1992; 84:771–6. [PubMed: 1573663]
9. Hu N, Wang C, Hu Y, Yang HH, Giffen C, Tang ZZ, Han XY, Goldstein AM, Emmert-Buck MR, Buetow KH, Taylor PR, Lee MP. Genome-wide association study in esophageal cancer using the genechip mapping 10K array. Cancer Res. 2005; 65:2542–6. [PubMed: 15805246]
10. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science. 1988; 241:1077–80. [PubMed: 3413476]
11. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005; 37:1217–23. [PubMed: 16244653]
12. Poliakov A, Cotrina M, Wilkinson DG. Diverse roles of Eph receptors and ephrins in the regulation of cell migration and tissue assembly. Dev Cell. 2004; 7:465–80. [PubMed: 15469835]
13. Katoh M, Katoh M. Comparative intergromics on Eph family. Int J Oncol. 2006; 28:1243–7. [PubMed: 16596241]
14. Batlle E, Bacani J, Begthel H, Jonkheer S, Gregorieff A, Van de Born M, Malats N, Sancho E, Boon E, Pawson T, Gallinger S, Pals S, Clevers H. EphB receptor activity suppresses colorectal cancer progression. Nature. 2005; 435:1126–30. [PubMed: 15973414]
15. Kojima T, Chang JH, Azar DT. Proangiogenic role of ephrinB1/EphB1 in basic fibroblast growth factor-induced corneal angiogenesis. Am J Pathol. 2007; 170:764–73. [PubMed: 17255342]
16. Liu C, Xu Z, Gupta D, Dziarski R. Peptidoglycan recognition proteins: a novel family of four human innate immunity pattern recognition molecules. J Biol Chem. 2001; 276:34686–94. [PubMed: 11461926]
17. Wang ZM, Li X, Cocklin RR, Wang M, Wang M, Fukase K, Inamura S, Kusumoto S, Gupta D, Dziarski R. Human peptidoglycan protein-L is an N-Acetylmuramoyl-L-alanine amidase. J Biol Chem. 2003; 278:49044–52. [PubMed: 14506276]
18. Xi L, Wang S, Wang H, Gupta D. Differential expression of peptidoglycan recognition protein 2 in the skin and liver requires different transcription factors. J Biol Chem. 2006; 281:20738–48. [PubMed: 16714290]
19. Bertucci F, Salas S, Eysteries S, Nasser V, Finetti P, Ginestier C, Charafe-Jauffret E, Loriod B, Bachelart L, Montfort J, Victorero G, Viret F, et al. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene. 2004; 23:1377–91. [PubMed: 14973550]
20. Yamada S, Ohira M, Horie H, Ando K, Takayasu H, Suzuki Y, Sugano S, Hirata T, Goto T, Matsunaga T, Hiyama E, Hayashi Y, et al. Expression profiling and differential screening between hepatoblastomas and the corresponding normal livers: identification of high expression of the

PLK1 oncogene as a poor-prognostic indicator of hepatoblastomas. Oncogene. 2004; 23:5901–11. [PubMed: 15221005]

21. Volinia S, Dhand R, Vanhaesebroeck B, MacDougall LK, Stein R, Zvelebil MJ, Domin J, Panaretou C, Waterfield MD. A human phosphatidylinositol 3-kinase complex related to the yeast Vps34p-Vps15p protein sorting system. EMBO J. 1995; 14:3339–48. [PubMed: 7628435]
22. Carter CJ. Multiple genes and factors associated with bipolar disorder converge on growth factor and stress activated kinase pathways controlling translation initiation: implications for oligodendrocyte viability. Neurochem Int. 2007; 50:461–90. [PubMed: 17239488]
23. Toker A, Yoeli-Lerner M. Akt signaling and cancer: surviving but not moving on. Cancer Res. 2006; 66:3963–6. [PubMed: 16618711]
24. Gandemer V, Rio AG, de Tayrac M, Sibut V, Mottier S, Ly Sunnaram B, Henry C, Monnier A, Berthou C, Le Gall E, Le Treut A, Schmitt C, et al. Five distinct biological processes and 14 differentially expressed genes characterize TEL/AML1-positive leukemia. BMC Genomics. 2007; 8:385–99. [PubMed: 17956600]
25. Orlowski J, Grinstein S. Diversity of the mammalian sodium/proton exchanger SLC9 gene family. Pflugers Ach. 2004; 447:549–65.
26. Harguindey S, Orive G, Pedraz JL, Paradiso A, Reshkin SJ. The role of pH dynamics and the Na^+/H^+ antiporter in the etiopathogenesis and treatment of cancer. Two faces of the same coin-one single nature. Biochim Biophys Acta. 2005; 1756:1–24. [PubMed: 16099110]
27. Tamura K, Furihata M, Tsunoda T, Ashida S, Takata R, Obara W, Yoshioka H, Daigo Y, Nasu Y, Kumon H, Konaka H, Namiki M, et al. Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles. Cancer Res. 2007; 67:5117–25. [PubMed: 17545589]

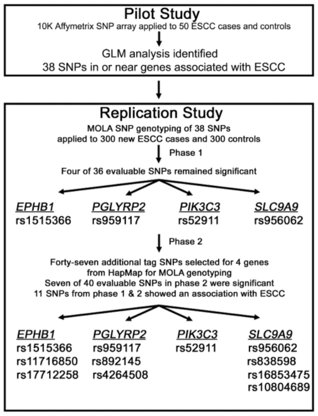
Pilot Study
10K Affymetrix SNP array applied to 50 ESCC cases and controls
GLM analysis identified
38 SNPs in or near genes associated with ESCC
Replication Study
MOLA SNP genotyping of 38 SNPs
applied to 300 new ESCC cases and 300 controls
Phase 1
Four of 36 evaluable SNPs remained significant
EPHB1
rs1515366
PGLYRP2
rs959117
PIK3C3
rs52911
SLC9A9
rs956062
Phase 2
Forty-seven additional tag SNPs selected for 4 genes
from HapMap for MOLA genotyping
Seven of 40 evaluable SNPs in phase 2 were significant
11 SNPs from phase 1 & 2 showed an association with ESCC
EPHB1
rs1515366
rs11716850
rs17712258
PGLYRP2
rs959117
rs892145
rs4264508
PIK3C3
rs52911
SLC9A9
rs956062
rs838598
rs16853475
rs10804689

Figure 1.

1. ESCC = Esophageal squamous cell carcinoma.
2. MOLA = multiplex oligonucleotide ligation assay.
3. *EPHB1* = Ephrin receptor EphB1.
4. *PGLYRP2* = Peptidoglycan recognition protein 2.
5. *PIK3C3* = Phosphatidylinositol 3-kinase, class 3.
6. *SLC9A9* = Solute carrier family 9, isoform A9.

Table I

CHARACTERISTICS AND ODDS RATIOS FOR ESCC CASES AND CONTROLS

Characteristics	Cases (n=300)	Controls (n=300)	OR[1]	95%CI	*p* value
Age (median, range, yrs)	59 (33-72)	59 (33-76)	1	1.0-1.0	0.99
Sex (% male)	100%	100%	NA	NA	NA
Tobacco use (% ever)	86%	88%	0.86	0.53-1.39	0.54
Alcohol use (% any)	72%	71%	1.05	0.74-1.49	0.79
Pickled vegetable use (ever)	51%	56%	0.83	0.60-1.14	0.25
Family history of UGI cancer	28%	18%	1.84	1.25-2.72	0.002

95%CI = 95% confidence interval.

ESCC = esophageal squamous cell carcinoma.

NA = not applicable.

OR = odds ratio.

UGI = upper gastrointestinal.

[1] Crude Odds Ratio.

Table II

ODDS RATIOS FOR SIGNIFICANT SNPs IN ESCC CASE-CONTROL STUDY (CASES n=300, CONTROLS n=300)

Gene	Location	SNP ID	common/variant allele frequency	OR/variant allele (95%) GLM	p^4 (GLM)	p^5 (2×3)	p^6 (HWE)
EPHB1	Intron 1	rs1515366[2]	T/A (0.77/0.23)	0.92 (0.86-0.99)	0.019	0.049	1.00
EPHB1	Intron 13	rs11716850[3]	C/T (0.70/0.30)	1.00 (0.94-1.07)	0.899	0.033	0.79
EPHB1	3' region	rs17712258[3]	T/A (0.70/0.30)	1.01 (0.94-1.07)	0.848	0.012	0.49
PGLYRP2	Intron 1	rs959117[2]	A/G (0.83/0.17)	0.93 (0.86-1.01)	0.061	0.181	0.55
PGLYRP2	Exon 2	rs892145[1]	T/A (0.64/0.36)	1.06 (1.00-1.12)	0.049	0.079	0.16
PGLYRP2	Intron 2	rs4264508	C/T (0.55/0.45)	0.94 (0.89-0.99)	0.017	0.033	0.02
PIK3C3	Intron 4	rs52911[2]	T/C (0.60/0.40)	0.93 (0.88-0.99)	0.020	0.015	0.54
SLC9A9	Intron 5	rs956062[2]	C/T (0.57/0.43)	1.03 (0.98-1.09)	0.222	0.024	1.00
SLC9A9	Intron 10	rs838598	A/G (0.60/0.40)	0.96 (0.90-1.01)	0.134	0.038	0.12
SLC9A9	Intron 14	rs16853475	C/T (0.78/0.22)	0.91 (0.85-0.98)	0.012	0.039	1.00
SLC9A9	Intron 14	rs10804689	G/A (0.76/0.24)	1.08 (1.01-1.15)	0.016	0.056	0.27

2×3 = 2×3 frequency distribution table.

ESCC = esophageal squamous cell carcinoma.

GLM = general linear model (additive model).

HWE = Hardy-Weinburg equilibrium.

SNP = single nucleotide polymorphism.

Homozygous common allele in controls is the reference genotype.

[1] SNP coding a nonsynonymous amino acid substitution Lysine 270 to Methionine.

[2] SNPs identified from pilot study.

[3] SNPs in linkage disequilibrium.

[4] Permuted *p* value GLM (additive model).

[5] Permuted Fishers *p* value (2×3).

[6] *P* value (HWE) derived from 2×3 distributions.

Table III

ODDS RATIOS FOR NUMBER OF PROTECTIVE ALLELES IN ESCC CASE-CONTROL STUDY (CASES n=300, CONTROLS n=300)

Number of protective alleles	Allele frequency (controls/cases)	OR[1] (95%CI) GLM	*p*[2] (GLM)
0-1 (ref.)	0.21/0.35	1.0 (ref.)	1.0
2	0.25/0.27	0.90 (0.80-1.01)	0.075
3	0.27/0.26	0.87 (0.78-0.98)	0.019
4-6	0.28/0.13	0.74 (0.66-0.84)	$1.53E^{-06}$
Test for trend across categories	NA	0.91 (0.88-0.95)	$2.86E^{-06}$

SNPs and their corresponding alleles used for calculating combined effects of multiple protective alleles; rs1515366 (allele A), *EPHB1*; rs4264508 (allele T), *PGLYRP2*; rs52911 (allele C), *PIK3C3*; rs16853475 (allele T), *SLC9A9*. The single most protective SNP from each of our four targeted genes were evaluated to determine the combined effects of multiple protective alleles.Multiple protective alleles were examined in three ways: (i) as a continuous variable; (ii) categorized into four groups approximating quartiles as: group 1 (0-1 protective allele), group 2 (2 protective alleles), group 3 (3 protective alleles), and group 4 (4-6 protective alleles; there were no individuals with 7-8 protective alleles) and analyzed using indicator variables; and (iii) combined into a single ordinal variable to test for trend.

[1] Odds ratio GLM additive mode.

[2] *P* value GLM additive model.