Exome Sequencing for Head and Neck Cancer Predisposition Genes

Yao Yu,[1*] Bingjian Feng,[2*] Chun-Pin Chang,[3] Russell Bell,[4] Austin Wood,[4] Erich Sturgis,[5] Guojun Li,[6] Andrew Olshan,[7] Chien-Jen Chen,[8] Pen-Jen Lou,[9] Wan-Lun Hsu,[10] Melissa Cessna,[11] Benjamin Witt,[12] Deb Neklason,[13] Mia Hashibe,[14] Chad Huff,[1] Sean Tavtigian[4]

1. Department of Epidemiology, Division of Cancer Prevention and Population Sciences, MD Anderson Cancer Center, Houston, TX
2. Department of Dermatology, University of Utah School of Medicine, Salt Lake City, UT
3. Division of Public Health, Department of Family and Preventive Medicine, University of Utah School of Medicine, Salt Lake City, UT
4. Department of Oncological Sciences, University of Utah School of Medicine, Salt Lake City, UT
5. Department of Otolaryngology - Head & Neck Surgery, Baylor College of Medicine, Houston, TX
6. Department of Head and Neck Surgery, Division of Surgery, MD Anderson Cancer Center, Houston, TX
7. Department of Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC
8. Genomic Research Center, Academia Sinica, Taipei, Taiwan
9. Department of Otolaryngology, National Taiwan University Hospital, Taipei, Taiwan
10. Master Program of Big Data in Medical Healthcare Industry, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan, and Data Science Center, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan
11. Intermountain Healthcare, Salt Lake City, UT
12. Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT
13. Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT
14. Department of Epidemiology, UCLA Fielding School of Public Health, Los Angeles, CA

*co-first authors

Corresponding author
Sean Tavtigian, PhD
sean.tavtigian@hci.utah.edu

Acknowledgements
Research reported in this publication was supported by the NIDCR of the NIH under award number R01DE023414 (Hashibe/Tavtigian). We thank the UPDB staff of Huntsman Cancer Institute, University of Utah (funded in part by the Huntsman Cancer Foundation) for their role in the ongoing collection, maintenance and support of the Utah Population Database (UPDB). We also acknowledge partial support for the UPDB through grant P30 CA2014 from the National Cancer Institute, University of Utah and from the University of Utah's program in Personalized Health and Utah Clinical and Translational Science Institute. The Utah Cancer Registry is funded by the National Cancer Institute's SEER Program, Contract No. HHSN261201800016I, the US Centers for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP007131, with additional support from the University of Utah and Huntsman Cancer Foundation.

Abstract

Introduction. While common variants have been studied for head and neck cancer (HNC) risk and exome sequencing has been conducted for head and neck tumor tissue samples, large-scale studies of exome sequencing on head and neck cancer risk have not been conducted. Our study aimed to identify head and neck cancer predisposition genes with exome sequencing and to assess interactions between the head and neck cancer susceptibility genes and tobacco use.

Methods. We conducted a case-control analysis to identify germline susceptibility genes for HNC risk. We used whole-exome sequencing data in Phase 1, targeted gene sequencing data in Phase 2 (2,134 HNC cases and 2,072 controls), and publicly available whole-exome sequencing data from the UK Biobank database. We conducted a gene-based association analysis and estimated the effect size of known and novel HNC susceptibility genes and pathways.

Results. Six known cancer predisposition genes reached nominal significance ($p < 0.05$), including *BRCA1*, *BRCA2*, *RAD51B*, *BAP1*, *APC*, and *MUTYH*. Additionally, we identified 21 novel HNC risk genes with meta p-value < 0.05. LoF variants in *BRCA1* (OR=5.0, 95% CI: 1.72-14.57), *BRCA2* (OR=2.56, 95% CI: 1.22-5.41), and *MUTYH* (OR=4.84, 95% CI: 1.01-13.86) exhibited significantly elevated effect sizes. *APC*, *BRCA2*, and the DNA repair gene set exhibited borderline interactions with smoking, with synergy indices and 95% confidence intervals of 1.92 [0.98, 3.75], 1.77 [0.91, 3.46], and 1.57 [0.97, 2.54], respectively.

Conclusions. We observed associations with HNC risk for DNA repair pathway genes associated with breast cancer and polyposis colorectal cancer predisposition genes. Potential interactions between these genes and tobacco smoking were also identified. Our study utilized targeted gene sequencing and whole-exome sequencing approaches to identify cancer susceptibility genes, while further studies are needed to confirm the associations observed with HNC risk.

Introduction

Each year in the world, approximately 771,000 head and neck cancer cases are diagnosed in the world (Ferlay 2024). Head and neck cancer includes cancer subsites of the oral cavity, oropharynx, hypopharynx, and larynx. The highest burden in the world of incident cases of head and neck cancer is in India (223,000 cases), China (84,000 cases) and the US (56,000 cases; Ferlay 2024). High age-standardized incidence rates of head and neck cancer are observed in Russia (9.9 per 100,000), the UK (9.8 per 100,000), and the United States (9.8 per 100,000; Ferlay 2024). Tobacco use and alcohol drinking are established risk factors with attributable risks ranging from 51% for North America, 83% for South America, and 84% for Europe (Hashibe 2009). HPV infections are a strong risk factor for oropharyngeal cancer (IARC 2012).

Previous genetic susceptibility studies on head and neck cancer largely focused on common genetic variants involved in the alcohol and tobacco metabolism and DNA repair pathways. A GWAS on European head and neck cancer patients reported associations with variants in the *ADH* gene cluster, near *ALDH2* and near the DNA repair gene *HEL308* (McKay 2011). In the Million Veteran Program Cohort including 4,012 HNC cases and 16,048 controls, variants associated with HNC risk were identified: loci in *DMBT1* and *TPM1* were associated with oral cavity, hypopharyngeal and laryngeal cancer while loci in *CRTAM*, *CHD5* and *CDH2* were associated with oropharyngeal cancer (Liu 2022). Another GWAS in 2,171 non-Hispanic white head and neck cancer cases and 4,493 controls reported on associations with *HLA-DQB1*, *CCHCR1*, and a novel locus 6p22.1.

Several GWAs or large-scale genotyping studies have been conducted in Asian and African-American populations. A GWAS study on 993 Chinese laryngeal cancer patients and 1,995 controls, with replication in 2,398 cases and 2,804 controls reported on three loci: *FADS1*, *AIF1*, and *TBX5* (Wei 2014). Another GWAS study on 1,529 Taiwanese oral cancer patients and 44,572 controls reported associations with variants in the *HLA-B* and *HLA-DQ* gene cluster, as well as in *TERT-CLMPT1L*, *ADH1B*, *LAMC3*, *SKIV2L* and *TNXB* genes (Bau 2023). In the Million Veteran Program Cohort, variants associated with HNC risk for African-American patients were identified in immune genes such as *HLA-G*, *GA1*, *CD6*, *NCAM1*/*CD56*, *CD68*, *SOCS6* (Liu 2022).

While common variants have been studied for head and neck cancer risk and exome sequencing has been conducted for head and neck tumor tissue samples, large-scale studies of exome sequencing on head and neck cancer risk have not been conducted to our knowledge. Our study aims to apply case-control exome sequencing to identify head and neck cancer predisposition genes and to assess whether there are interactions between the head and neck cancer susceptibility genes and tobacco use.

Results

Phase 1 included 220 familial HNC cases, while Phase 2 was comprised of 2,134 HNC cases and 2,072 controls (Table 1). In both phases, a high proportion of cases were diagnosed with oropharyngeal cancer (39.5% in Phase 1 and 29.8% in Phase 2). More than 70% of cases and controls were males. Phase 2 included a higher number of Asian and Pacific Islander cases compared to Phase 1, reflecting the recruitment of a large proportion of participants from the Taiwan center. As expected, the percentage of individuals who were ever cigarette smokers or alcohol drinkers was higher among cases in both Phase 1 and Phase 2, compared to controls in Phase 2.

We conducted whole-exome sequencing using Phase 1 samples with a targeted depth of 60X for each sample. We conducted joint genotype calling using the sequencing data of recruited samples together with samples from the 1000 Genome project. We then performed gene-based association tests, gene prioritization combining gene-based burden test results and biological relevant scores, and gene set-based association tests using VICTOR and VAAST (Hu 2013) separately. We consolidated the top 100 genes from each of the six tests into a gene panel of known or suspected cancer susceptibility genes. As a result, we obtained a list of 493 genes for HNC risk and designed a targeted sequencing panel.

Targeted gene sequencing, designed to achieve high coverage and precision for specific genomic regions, offers a cost-effective strategy for genetic association studies. To investigate genetic susceptibility to HNC, we performed targeted sequencing on 2,134 HNC cases and 2,072 controls recruited in Phase 2, followed by a gene-based association analysis. Adhering to GATK's best practices for joint genotype calling, we implemented rigorous quality control measures to ensure high data quality, retaining only consistent and high-confidence genotypes for analysis. This process identified 175,277 single nucleotide variants (SNVs) and 5,518 insertions/deletions (INDELs) with minor allele frequencies (MAF) < 0.005. Of these, 34,308 variants were located in protein-coding or splicing regions. Among the exonic variants, 1,004 were Loss-of-function (LoF) variants (including nonsense, frameshift INDELs, and splicing-altering variants), 373 were in-frameshift variants, and 19,887 were missense variants.

In case-control association analyses, aggregating multiple variants in a single test is critical for effectively studying rare variants. We conducted gene-based association tests on 493 targeted genes using VAAST2 across four populations: non-Hispanic whites (NHW), African Americans (AA), Hispanics (HIS), and Asians (ASI). VAAST2 employs an advanced variant prioritization framework that integrates functional impact and phylogenetic conservation data, enabling precise differentiation between benign and potentially pathogenic variants and improving the accuracy and relevance of genetic association results. To enhance the power of our gene-based tests, we also performed two additional logistic regression model-based tests, focusing on LoF variants and predicted damaging missense variants. We then integrated VAAST results and the two logistic regression tests into an omnibus test using the aggregated Cauchy association test (ACAT) (Figure 1). To account for analyses across the four populations, we conducted a meta-

analysis, combining ACAT p-values and weighting them by the effective sample size of each population. In the multi-population meta-analysis, we observed nominally significant associations in several known cancer predisposition genes, including *BRCA1* (p = 0. 0462), *BRCA2* (p=0.0198), *RAD51B* (p = 0.006), *BAP1* (p = 0.0268), *APC* (p=0.0192), and *MUTYH* (p=0.0371) (Table 2).

To validate our findings using targeted gene sequencing data from Phase 2 samples, we extended the association analysis to include HNC cases and cancer-free controls from the UK Biobank. We obtained whole-exome sequencing data from 2,659 cases and 121,788 controls and performed joint genotype calling using DeepVariant implemented in GLnexus, followed by quality control using XPAT2. Utilizing a similar approach to the one described earlier, we conducted a whole-exome-wide omnibus test that incorporated analyses of all variants with prioritization scores from VAAST2, LoF variants, and predicted damaging variants for NHW population. We observed no inflation in ACAT p values in the association tests using UK Biobank samples. Although no gene reached genome-wide significance, we successfully replicated two known cancer predisposition genes: *BRCA1* (p = 0. 0494) and *BRCA2* (p = 0.0284).

We then conducted a meta-analysis to combine the analyses using the 3-center data and UK Biobank data for the 493 targeted genes. Six known cancer predisposition genes reached a nominal significance ($p < 0.05$), including *BRCA1*, *BRCA2*, *RAD51B*, *BAP1*, *APC*, and *MUTYH*. Additionally, we identified 21 novel HNC risk genes with meta p-value < 0.05 (Table 3). Using UK Biobank data exclusively, we also observed several genes associated with HNC risk under a suggestive significance level of 0.001, which were not part of the targeted sequencing panel, including *PCDHB5*, *TANC1*, and *ADGRA2* (Table 4).

To characterize the contribution of rare genetic variation to the risk of HNC, we evaluated the effect sizes of LoF variants and predicted damaging variants in NHW (Figure 2). The LoF variants in *BRCA1* (OR=5.0, 95% CI: 1.72-14.57), *BRCA2* (OR=2.56, 95% CI: 1.22-5.41), and *MUTYH* (OR=4.84, 95% CI: 1.01-13.86) exhibited significantly elevated effect sizes. The LoF variants in *APC*, BAP1, and *RAD51B* also showed large odds ratios (5.36, 11.95, and 2.11, respectively), though with wide confidence intervals due to small sample sizes. In contrast, we found missense variants in these established HNC genes only confer a moderate effect size. Furthermore, LoF variants in other top genes from UK Biobank data, including *SLC41A2*, *STAT1*, *ADAM12*, *KRT5*, *NPAT*, and *SH3D21*, also showed large effect sizes for HNC risk (Figure S3). Interestingly, predicted damaging missense variants in *RFX5*, *CLTCL1*, *KLF4*, *XRCC6*, *LEMD3*, and *AK2* exhibit significantly large effect sizes for HNC (Figure S5).

In addition to gene-based rare variant association tests, we performed gene set analysis (GSA) to assess the joint effects of biologically correlated genes. We conducted a pathway-based analysis using Fisher's method to combine the gene-based meta p values in selected pathways. We observed that the pathways of DNA homologous recombination repair core susceptibility genes, HGNC: 548, and homologous recombination repair of replication independent double strand

breaks from REACTOME all have nominal significant p values (Table 5). We observed no evidence for the association from three Gene Ontology (GO) pathways (GO: negative regulation of endothelial cell migration, GO: stem cell division, and GO: regulation of fibroblast migration) as well as the pathway of DNA mismatch repair core susceptibility genes (negative control).

We further conducted a pathway-based effect size estimation, combining LoF and predicted damaging variants for HNC risk. Separate analyses by population were followed by a meta-analysis which further supported the effect size of these pathways. The DNA homologous recombination repair pathway has an OR of 1.67 [95% CI: 1.26-2.22] with both LoF and predicted damaging variants and an OR of 1.64 [95% CI: 1.20-2.25] with LoF variants alone. We observed no evidence of association in other pathways including negative regulation of endothelial cell migration, stem cell division, and regulation of fibroblast migration. (Table S1 to S3)

We used the Synergy Index (S) to investigate whether the HNC susceptibility genes had an interaction with life-style risk factors such as smoking and drinking. *APC*, *BRCA2*, and the DNA repair gene set exhibited borderline interactions with smoking, with synergy indices and 95% confidence intervals of 1.92 [0.98, 3.75], 1.77 [0.91, 3.46], and 1.57 [0.97, 2.54], respectively.

Discussion

We identified multiple DNA homologous repair pathway genes where loss of function mutations had the biggest effect sizes including *BRCA1*, *BRCA2*, and *RAD51B*. In addition, we identified associations in the cancer predisposition genes, in *APC*, *BAP1*, and *MUTYH*. *MUTYH* is a recessive cancer predisposition gene leading to colonic adenomatous polyposis and cancer, but there is some speculation that a single altered copy slightly increases risk. *MUTYH* is involved in oxidative damage repair, which is associated with alcohol and smoking exposures.

Hypermethylation resulting in inactivation of *APC* gene expression resulting in tumorigenesis is found in oral cancers (Irimie 2018). If one copy of *APC* bears a germline pathogenic variant, and the other copy of *APC* becomes hypermethylated due to smoking, alcohol, or chronic inflammation, tumorigenesis would be triggered. Based on our results, potentially pathogenic variants in the *APC* gene have the largest effect on oral cancers (Table S7), and there is evidence of an interaction with smoking (Table 6), consistent with a hypothesis that this is a likely mechanism.

Smoking and or alcohol can cause DNA double-strand breaks, oxidative damage, and hypermethylation leading to the loss of expression of genes important in maintaining cell health. One insight from this study is that the identified genes are important in maintaining the health

and balance of cells, and when cells are subjected to damage from smoking and alcohol, their compromised ability to repair or control cell growth results in HNC cancer.

Some limitations of our study include the use of a heterogeneous population, which can create some challenges especially when investigating rare variants. Although our sample size was fairly large, some of the race and ethnicity groups had smaller numbers, which led to less statistical power to detect associations. In this study, we were able to navigate this challenge by considering the combined effect of genes (not variants) on HNC risk, matching controls appropriately, and pulling the different ethnic groups together through a meta-analysis.

The strengths of our study include the collaboration across multiple sites in the US and the international collaboration with Taiwan. The familial design in phase I as well as our gene-based approach contributed to an enhanced statistical power to find new head and neck cancer susceptibility genes. Our analytic approach was rigorous in combining the data across different race and ethnicity groups. We also had two parallel analyses in Utah and Texas, which acted as quality control for our analytic approach.

Conclusion

We observed associations with head and neck cancer risk for DNA repair pathway genes associated with breast cancer and colorectal cancer predisposition genes. Potential interactions between these genes and tobacco smoking were also identified. Thus, the most important result may be that carriers of germline pathogenic variants in DNA double strand break repair pathways genes should avoid tobacco smoking. Further studies are needed to confirm the associations observed with head and neck cancer, but our study supports our exome sequencing approach to identify cancer susceptibility genes.

References

Bau DT, Liu TY, Tsai CW, Chang WS, Gu J, Yang JS, Shih LC, Tsai FJ. A Genome-Wide Association Study Identified Novel Genetic Susceptibility Loci for Oral Cancer in Taiwan. Int J Mol Sci. 2023 Feb 1;24(3):2789. doi: 10.3390/ijms24032789. PMID: 36769103; PMCID: PMC9917812.

Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2024). Global Cancer Observatory: Cancer Today (version 1.1). Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today, accessed [12 November 2024].

Hashibe M, Brennan P, Chuang SC, Boccia S, Castellsague X, Chen C, Curado MP, Dal Maso L, Daudt AW, Fabianova E, Fernandez L, Wünsch-Filho V, Franceschi S, Hayes RB, Herrero R, Kelsey K, Koifman S, La Vecchia C, Lazarus P, Levi F, Lence JJ, Mates D, Matos E, Menezes A, McClean MD, Muscat J, Eluf-Neto J, Olshan AF, Purdue M, Rudnai P, Schwartz SM, Smith E, Sturgis EM, Szeszenia-Dabrowska N, Talamini R, Wei Q, Winn DM, Shangina O, Pilarska A, Zhang ZF, Ferro G, Berthiller J, Boffetta P. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. Cancer Epidemiol Biomarkers Prev. 2009 Feb;18(2):541-50. doi: 10.1158/1055-9965.EPI-08-0347. Epub 2009 Feb 3. PMID: 19190158; PMCID: PMC3051410.

Hu H, Huff C D, Moore B, et al. VAAST 2.0: Improved variant classification and disease□gene identification using a conservation□controlled amino acid substitution matrix[J]. Genetic epidemiology, 2013, 37(6): 622-634.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Biological agents. IARC Monogr Eval Carcinog Risks Hum. 2012;100(Pt B):1-441. PMID: 23189750; PMCID: PMC4781184.

Irimie AI, Ciocan C, Gulei D, Mehterov N, Atanasov AG, Dudea D, Berindan-Neagoe I. Current Insights into Oral Cancer Epigenetics. Int J Mol Sci. 2018 Feb 27;19(3):670. doi: 10.3390/ijms19030670. PMID: 29495520; PMCID: PMC5877531.

Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet. 2019 Mar 7;104(3):410-421. doi: 10.1016/j.ajhg.2019.01.002. PMID: 30849328; PMCID: PMC6407498.

Liu Y, Kramer JR, Sandulache VC, Yu R, Li G, Chen L, Yusuf ZI, Shi Y, Pyarajan S, Tsavachidis S, Jiao L, Mierzwa ML, Chiao E, Mowery YM, Shuman A, Shete S, Sikora AG, White DL. Immunogenetic Determinants of Susceptibility to Head and Neck Cancer in the

Million Veteran Program Cohort. Cancer Res. 2023 Feb 3;83(3):386-397. doi:10.1158/0008-5472.CAN-22-1641. PMID: 36378845; PMCID: PMC9896026.

McKay JD, Truong T, Gaborieau V, et al. A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. PLoS Genet. 2011 Mar;7(3):e1001333. doi: 10.1371/journal.pgen.1001333. Epub 2011 Mar 17. Erratum in: PLoS Genet. 2011 Apr;7(4). doi: 10.1371/annotation/9952526f-2f1f-47f3-af0f-1a7cf6f0abc1. Voodern, Tõnu [corrected to Vooder, Tõnu]. PMID: 21437268; PMCID: PMC3060072.

Shete S, Liu H, Wang J, Yu R, Sturgis EM, Li G, Dahlstrom KR, Liu Z, Amos CI, Wei Q. A Genome-Wide Association Study Identifies Two Novel Susceptible Regions for Squamous Cell Carcinoma of the Head and Neck. Cancer Res. 2020 Jun 15;80(12):2451-2460. doi: 10.1158/0008-5472.CAN-19-2360. Epub 2020 Apr 10. PMID: 32276964; PMCID: PMC7299763.

Wei Q, Yu D, Liu M, Wang M, Zhao M, Liu M, Jia W, Ma H, Fang J, Xu W, Chen K, Xu Z, Wang J, Tian L, Yuan H, Chang J, Hu Z, Wei L, Huang Y, Han Y, Liu J, Han D, Shen H, Yang S, Zheng H, Ji Q, Li D, Tan W, Wu C, Lin D. Genome-wide association study identifies three susceptibility loci for laryngeal squamous cell carcinoma in the Chinese population. Nat Genet. 2014 Oct;46(10):1110-4. doi: 10.1038/ng.3090. Epub 2014 Sep 7. PMID: 25194280.

Table 1. Characteristics of familial head and neck cancer (HNC) cases in Phase 1 and HNC cases and controls in Phase 2 of the study

	Phase 1 (N=220)		Phase 2 (N=4,206)			
	Familial HNC cases		HNC Cases (N=2,134)		Control (N=2,072)	
	N	%	N	%	N	%
Study center						
Houston	96	43.6	680	31.9	781	37.7
North Carolina	63	28.6	1,053	49.3	914	44.1
Taiwan	13	5.9	401	18.8	377	18.2
Utah	48	21.8	NA		NA	
Cancer subsite						
Oral Cavity	63	28.6	514	24.1		
Oropharynx	87	39.5	635	29.8		
Hypopharynx	**		146	6.8		
Oral Cavity - Oropharynx-Hypopharynx, NOS	**		317	14.9		
Larynx	44	20.0	522	24.5		
Sex						
Female	60	27.3	447	20.9	534	25.8
Male	160	72.7	1,687	79.1	1,538	74.2
Age (years)						
<55	72	32.8	827	38.8	817	39.4
55 to <65	78	35.4	711	33.3	649	31.3
65+	70	31.8	596	28.0	606	29.2
Race/ethnicity						
Non-Hispanic White	175	79.5	1,421	66.6	1,414	68.2
Black/African American	24	10.9	259	12.1	208	10.0
Hispanic	**		30	1.4	57	2.8
Asian and Pacific Islanders	**		405	19.0	382	18.4
Cigarette smoking						
Never cigarette smoker	58	26.4	441	20.7	950	45.8
Ever cigarette smoker	152	69.1	1,690	79.2	1,122	54.2
Alcohol drinking						
Never alcohol drinkers	47	21.4	477	22.4	839	40.5
Ever alcohol drinkers	162	73.6	1,654	77.5	1,233	59.5

*Other included non-specific or unknown race

** suppressed since one of the cell counts was <11

Table 2. P-values for association with head and neck cancer risk for known cancer predisposition genes combining results from 3 centers (Houston, North Carolina, and Taiwan) and UK Biobank

Gene	Meta p-value	p from 3 centers	p from UKB
BRCA1	0.0124	0.0462	0.0494
RAD51B	0.0147	0.0060	0.5569
APC	0.0192	0.0658	0.0538
BRCA2	0.0198	0.0894	0.0284
BAP1	0.0268	0.0223	0.3817
MUTYH	0.0371	0.0875	0.1007

Table 3. P-values for association with head and neck cancer risk for genes with Meta p-value < 0.05 combining results from 3 centers (Houston, North Carolina, and Taiwan) and UK Biobank

Gene	Meta p-value	p from 3 centers	p from UKB
SLC41A2	0.0013	0.0303	0.0015
CEP76	0.0079	0.0150	0.1459
NUP54	0.0150	0.0060	0.5641
KLLN	0.0152	0.0203	0.2282
ZNF280D	0.0196	0.0026	0.8337
STAT1	0.0197	0.0872	0.0299
UAP1L1	0.0198	0.0680	0.0531
RFX5	0.0215	0.0548	0.0930
ADAM12	0.0215	0.0243	0.2808
CLTCL1	0.0221	0.0807	0.0451
CDH23	0.0258	0.0031	0.8643
ACACB	0.0260	0.0472	0.1591
KRT5	0.0323	0.0885	0.0768
CPTP	0.0342	0.0740	0.1191
NPAT	0.0368	0.3420	0.0008
KLF4	0.0380	0.0959	0.0881
FGFR4	0.0382	0.0749	0.1399
XRCC6	0.0389	0.3251	0.0013
SH3D21	0.0390	0.1139	0.0639
LEMD3	0.0453	0.1471	0.0465
AK2	0.0485	0.0637	0.2540

Table 4. P-values for association with head and neck cancer risk for genes not included in the Phase 2 targeted sequencing panel with p<0.001 from the UK Biobank

Gene	p from UKB*
PCDHB5	9.93E-06
TANC1	4.64E-05
ADGRA2	6.58E-05
KDM8	0.0002
LACTBL1	0.0002
B4GALNT2	0.0002
AGER	0.0003
CDC6	0.0007
FZD7	0.0007
CHTOP	0.0008
CDR2	0.0008
SLPI	0.0008

* Threshold for the suggestive significant p-value was calculated as 0.05/(19500/500)

Table 5. Pathway analysis for pathways with p<0.25 for the association with head and neck cancer risk

Pathway	Genes	Meta p-value
DNA homologous recombination repair core susceptibility genes	*BRCA1,BARD1,BRCA2,PALB2,RAD51B,RAD51C*	0.0004
HGNC 548: Fanconi anemia complementation groups	*BRCA1,BRCA2,BRIP1,ERCC4,FANCA,FANCB,FANCC,FANCD2,FANCE,FANCF,FANCG,FANCI,FANCL,FANCM,MAD2L2,PALB2,RAD51,RAD51C,SLX4,UBE2T,XRCC2,RFWD3*	0.0364
REACTOME: homologous recombination repair of replication independent double strand breaks	*RAD50,H2AX,LIG1,MRE11,NBN,ATM,RAD51,RAD52,RPA1,RPA2,RPA3,BRCA1,BRCA2,TP53BP1,BRIP1,MDC1*	0.0431
GO: negative regulation of endothelial cell migration	*THBS1,KLF4,DAB2IP,APOH,TGFB1,SP100,SLIT2,GDF2,ITGB1BP1,ANGPT2,CXCL13,STARD13,VASH1,ACVRL1,DLL4,PTPRM,PDCD10,AGTR2,RGCC,CSNK2B,DCN,ANGPT4,HMGB1,APOE,FGF2,SYNJ2BP,SERPINF1,STC1,MAP2K5,HDAC5,RHOA,SVBP,NOTCH1,NR2F2,HRG,BMP10,AGER,KRIT1,MMRN2*	0.0691
GO: stem cell division	*PAFAH1B1,FZD7,ZFP36L2,CDKN2A,WNT3A,TEAD3,NUMBL,TIAL1,TGFB2,FGF13,LEF1,WNT7A,DCT,NUMB,FGFR2,KIT,ASPM,WWTR1,VANGL2,ETV5,NOTCH1,HOXB4,ING2,CUL3,DOCK7,RAB10,FGFR1,SOX5,ZBTB16*	0.1552
GO: regulation of fibroblast migration	*BAG4,UTS2,PRR5L,RCC2,RAC1,GNA12,SDC4,AGER,DMTN,FER,WDPCP,DDR2,SLC8A1,HAS1,BRAF,TGFB1,RFFL,CYGB,ARHGAP4,THBS1,AKT1,ARHGEF7,HYAL2,PRKCE,FGF2,MTA2,ITGB1BP1,CORO1C*	0.2289
Negative control: DNA mismatch repair core susceptibility genes	*MLH1,MSH2,MSH6,PMS2*	0.9732

Table 6. Test for interaction (i.e. supra-multiplicative ORs) between cigarette smoking and pathway and genes on the risk of head and neck cancer using LOF and predicted damaging variants, combining results from 3 centers (Houston, North Carolina, and Taiwan) and UK Biobank*

Gene/pathway	OR for smoking	OR for genes	OR for both	p-value
APC	2.93[2.62,3.27]	1.08[0.53,2.23]	4.87[3.15,7.53]	0.0277
DNA homologous recombination repair core susceptibility genes	2.97[2.65,3.33]	1.73[1.14,2.63]	5.24[3.75,7.34]	0.03325
BRCA2	2.94[2.63,3.29]	1.14[0.62,2.08]	4.68[2.95,7.43]	0.04676
AK2	2.94[2.64,3.29]	1.29[0.25,6.63]	8.16[3.03,22.0]	0.06064
XRCC6	2.95[2.64,3.30]	6.35[1.26,31.9]	24.1[8.12,71.6]	0.10467
RAD51B	2.95[2.64,3.30]	2.17[0.51,9.31]	8.02[2.77,23.2]	0.15696
PALB2	2.95[2.64,3.30]	1.81[0.43,7.60]	6.97[2.11,23.1]	0.18449
BRCA1	2.95[2.64,3.30]	1.59[0.55,4.62]	5.05[2.54,10.1]	0.19876
NUP54	2.95[2.64,3.29]	1.26[0.14,11.6]	5.84[1.85,18.4]	0.2058
BAP1	2.95[2.64,3.30]	1.55[0.26,9.21]	6.10[1.36,27.3]	0.25311
FGFR4	2.95[2.64,3.30]	1.10[0.15,7.83]	4.66[1.48,14.6]	0.26299
CPTP	2.95[2.64,3.30]	2.06[0.47,8.96]	5.28[2.16,12.9]	0.32192
RAD51C	2.97[2.66,3.32]	5.17[1.32,20.2]	6.39[1.42,28.7]	0.54786
KLLN	2.96[2.65,3.30]	4.71[0.65,34.0]	5.63[0.78,40.7]	0.55437
BARD1	2.97[2.66,3.32]	2.89[0.94,8.87]	3.42[1.14,10.2]	0.6996
STAT1	2.96[2.65,3.31]	13.9[2.43,79.3]	7.15[0.71,72.4]	0.70861
ADAM12	2.96[2.65,3.31]	2.90[0.78,10.8]	3.41[1.40,8.32]	0.71942
KRT5	2.96[2.65,3.31]	3.10[0.87,11.0]	3.44[1.48,7.99]	0.7437
CLTCL1	2.97[2.66,3.32]	7.11[2.26,22.4]	3.08[0.43,22.1]	0.8056

*Adjusted for race and sex

Figure 1. Study design

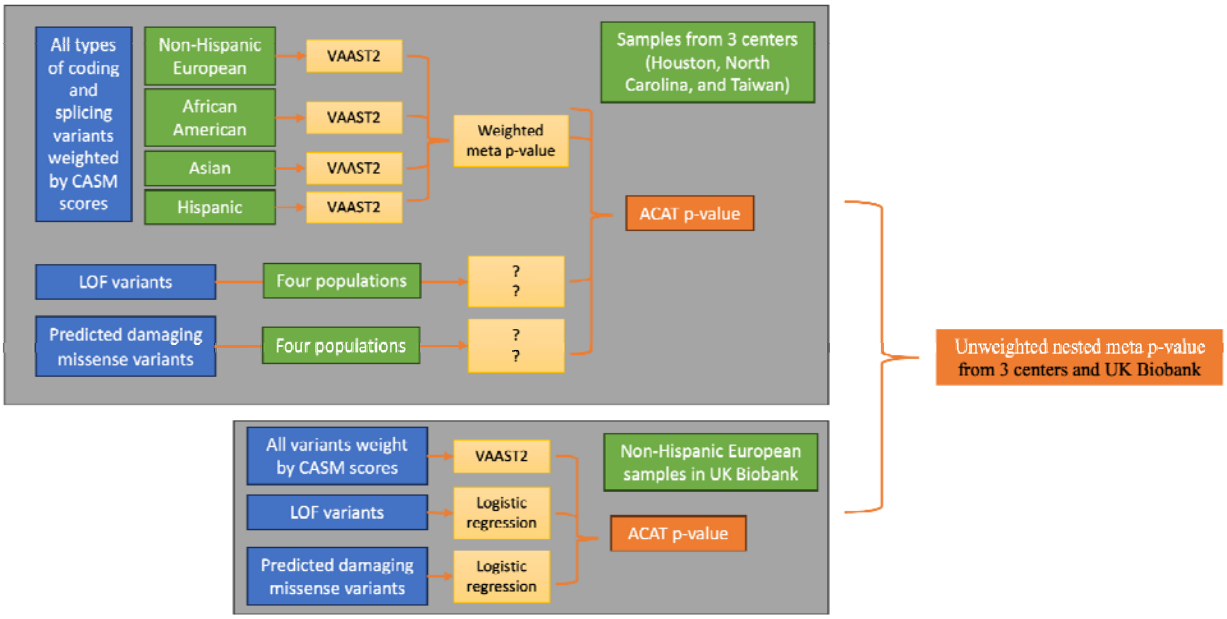


Figure 2. Forest plot of OR of various types of variants in known cancer predisposition genes (with evidence of association of cancers) after combining results from 3 centers (Houston, North Carolina, and Taiwan) and UK Biobank.

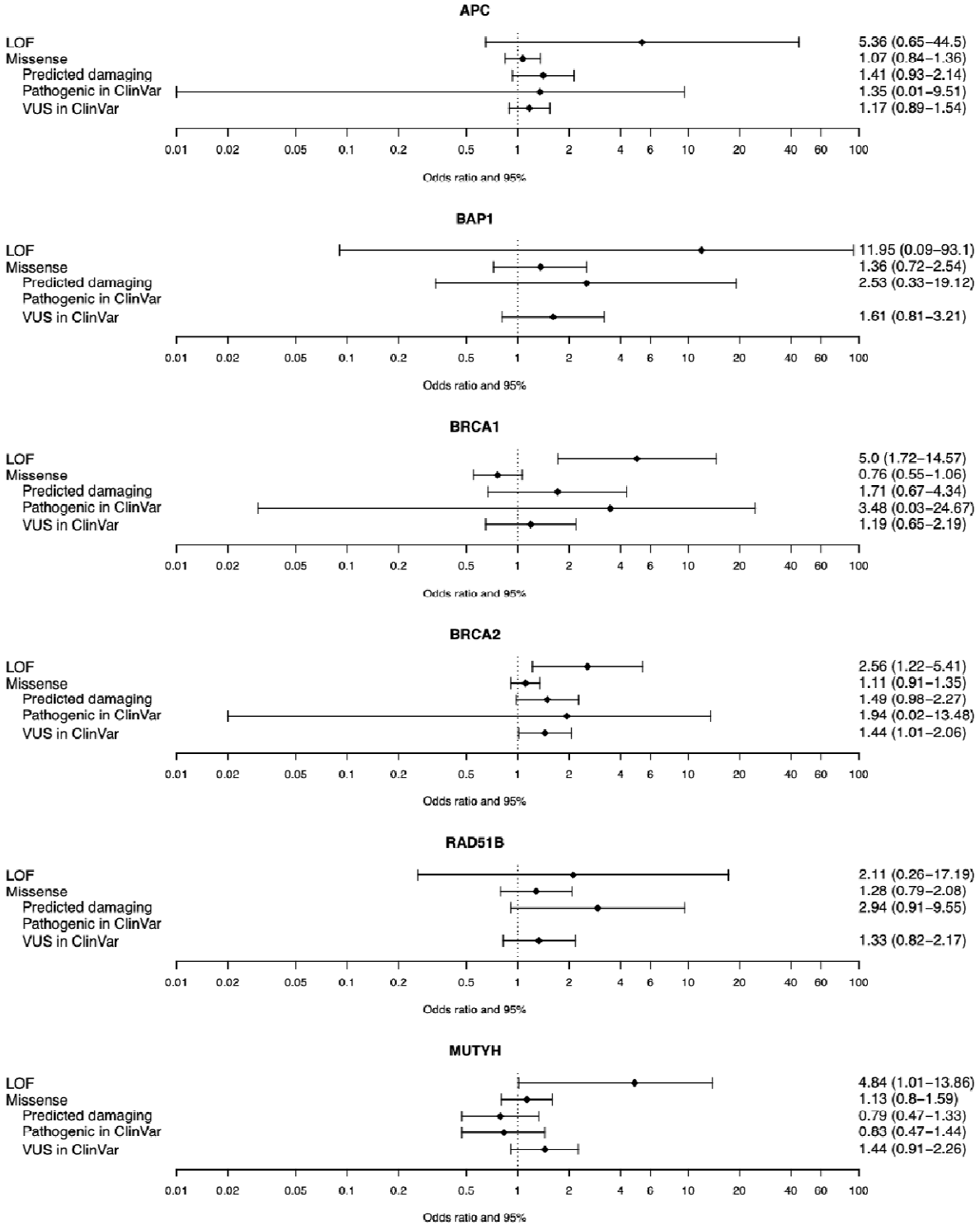


Figure S1. Q–Q plot of Meta P-value for gene-based association analysis for head and neck cancer risk.

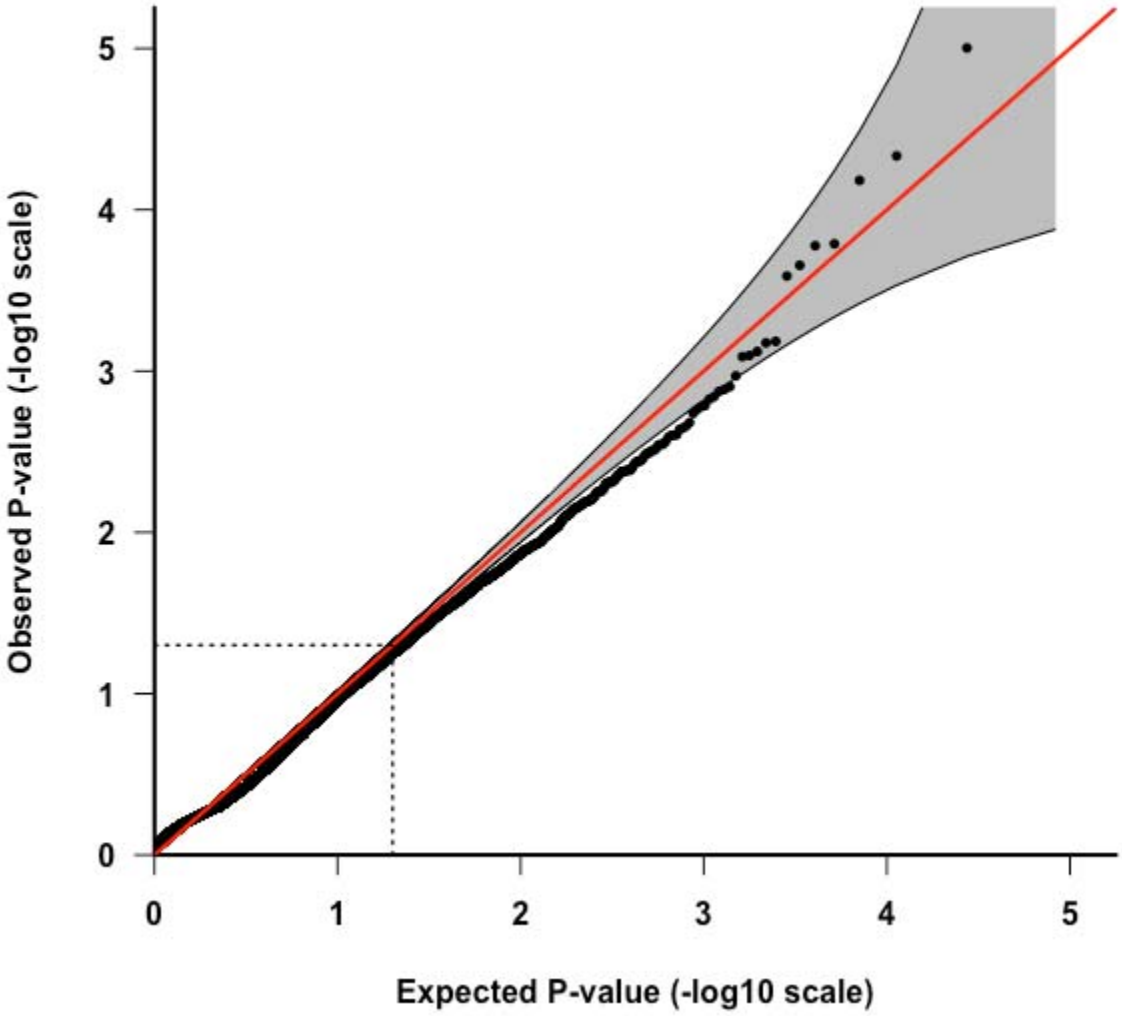


Figure S2. ORs of all variants in for genes with Meta p-value < 0.05 after combining results from Houston, North Carolina, Taiwan, and UK Biobank in non-Hispanic European population.

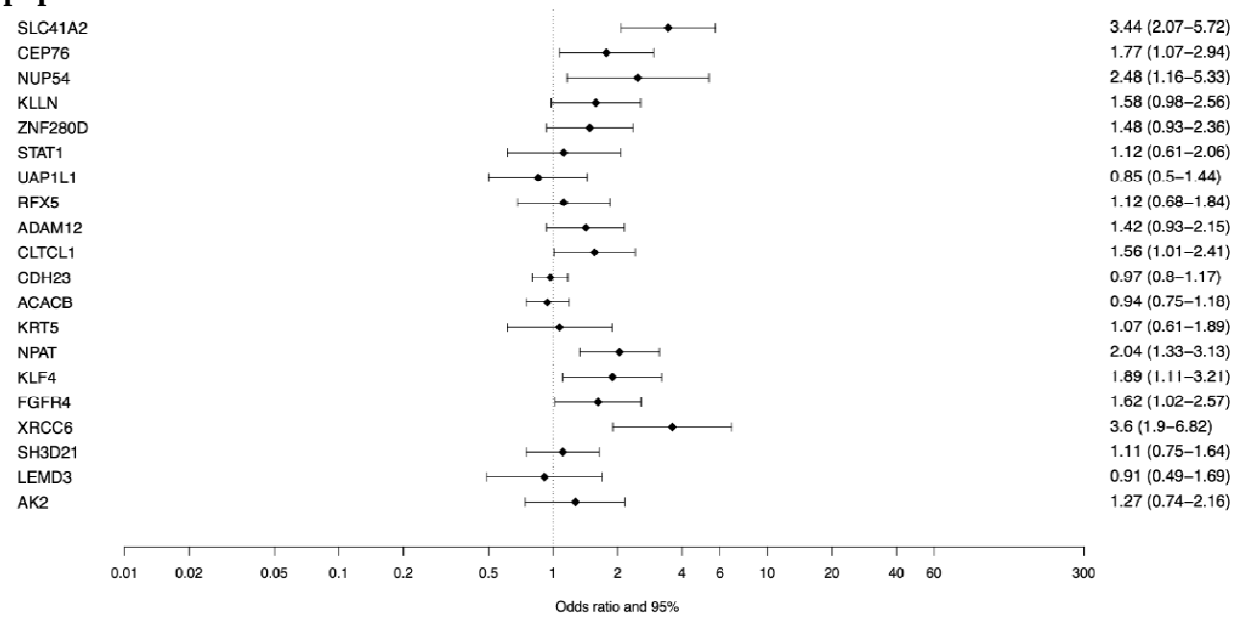


Figure S3. ORs of LOF variants in for genes with Meta p-value < 0.05 after combining results from Houston, North Carolina, Taiwan, and UK Biobank in non-Hispanic European population.

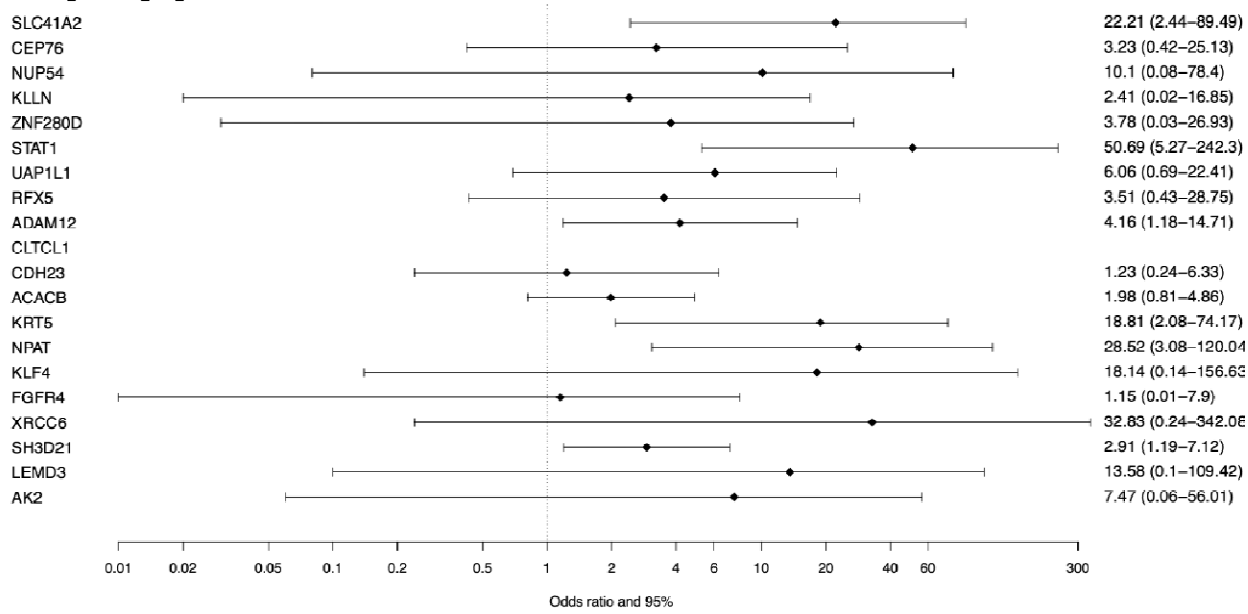


Figure S4. ORs of all missense variants in for genes with Meta p-value < 0.05 after combining results from Houston, North Carolina, Taiwan, and UK Biobank in non-Hispanic European population.

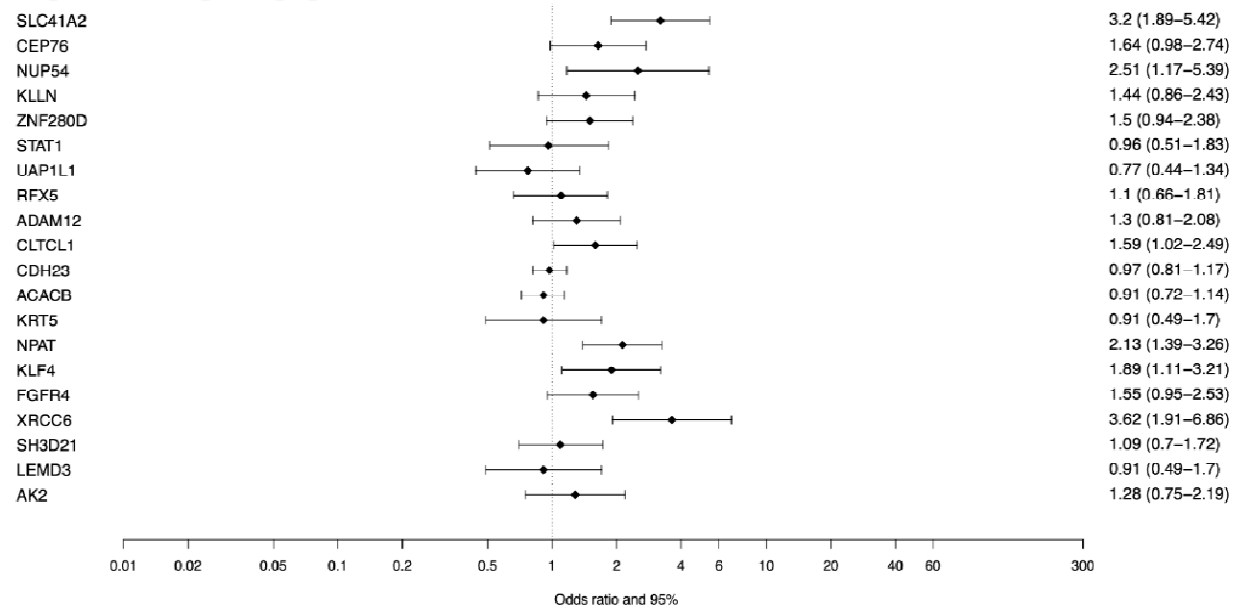
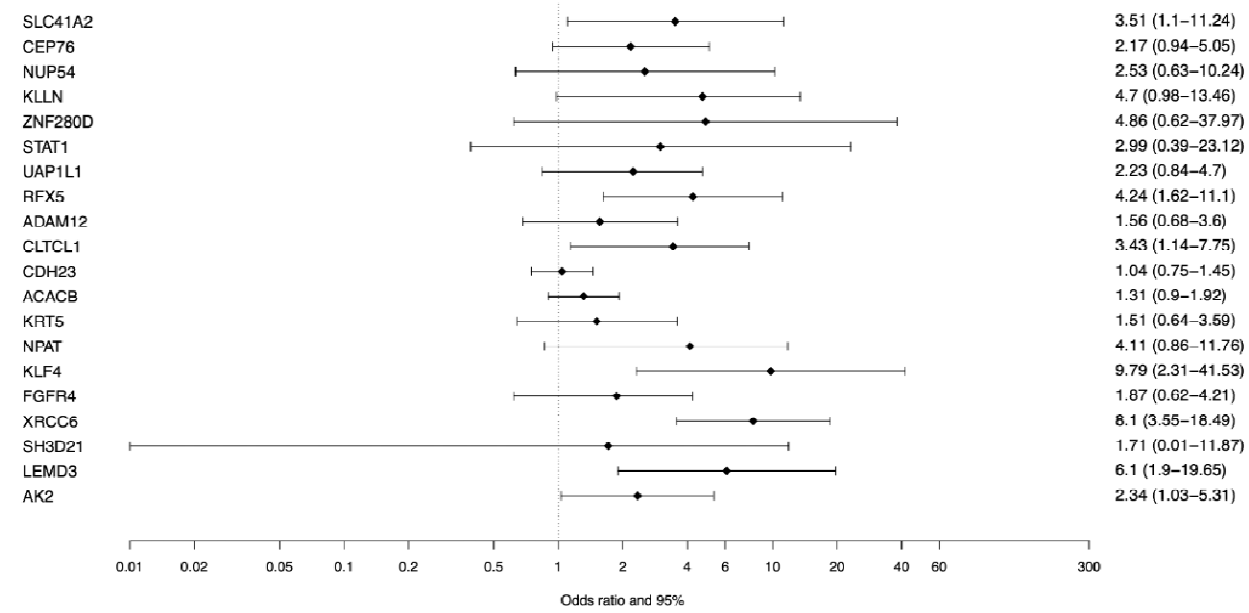


Figure S5. ORs of predicted pathogenic missense variants in for genes with Meta p-value < 0.05 after combining results from Houston, North Carolina, Taiwan, and UK Biobank in non-Hispanic European population.



Material and Methods

Study cohorts of head and neck cancer

In this study, we conducted a multi-center case-control analysis for HNC. HNC cancer subsites were defined by the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) codes. Codes for oral cavity were C00.3, C00.4, C00.5, C00.6, C00.8, C00.9, C02.0, C02.1, C02.2, C02.3, C03.0, C03.1, C03.9, C04.0, C04.1, C04.8, C04.9 C05.0, C06.0, C06.1, C06.2, C06.8, and C06.9. Codes for oropharynx were C01.9, C02.4, C05.1, C05.2, C09.0, C09.1, C09.8, C09.9, C10.0, C10.2, C10.3, C10.4, C10.8, and C10.9. Codes for hypopharynx were C12.9, C13.0, C13.1, C13.2, C13.8, and C13.9. Codes for larynx were C10.1, C32.0, C32.1, C32.2, C32.3, C32.8, C32.9. Codes for oral cavity-oropharynx-hypopharynx, not otherwise specified (NOS) were C02.8, C02.9, C05.8, C05.9, C14.0, C14.2, C14.8.

The sample recruitment of this study consisted of two phases. In Phase 1, we included only familial HNC cases from study centers in Houston, North Carolina, Taiwan, and Utah. In Phase 2, we included non-familial HNC cases and controls from the study centers in Houston, North Carolina, and Taiwan. Demographic characteristics were collected via questionnaires administered by trained interviewers at each study center. DNA was extracted from blood samples. In Phase 1, the final analysis included 220 familial HNC cases. In Phase 2, there were 2,134 HNC cases and 2,072 controls. The study was approved by Institutional Review Boards at each of the institutes involved. To perform a meta-analysis, we obtained whole exome sequencing data from UK Biobank (200K release), which included 531 HNC patients per the selection criteria above and 119,716 cancer-free controls.

Whole exome sequencing of Phase 1 samples

We first conducted whole exome sequencing (WES) using Phase 1 samples with an average depth of over 60x using Illumina HiSeq 2500 at the HCI High-Throughput Genomics Core Facility at the University of Utah Health Sciences. The FASTQ reads were mapped to the human genome reference GRCh37 downloaded from the GATK resource using BWA-MEM (v0.7.10) {Li 2009 19451168}). Duplicates were marked using Samblaster (v0.1.22){Faust 2014 24812344}, sorted, and formatted with Sambamba {Tarasov 2015 25697820}. We also obtained WES data of controls from the 1000 Genomes Project (G1K). The 1000 Genomes samples included non-Finnish Europeans (NFE), African (AFR), and East Asian (EAS) descendants to match with the cases. We conducted variant calling using GATK HaplotypeCaller to create genome VCF files, followed by joint genotype calling for all samples, and recalibrated the quality of variant calls following GATK's best practice.

We then conducted variant and sample level quality controls by using the following criteria: (1) Exclude genotypes with read depth <10x or genotype quality <55. (2) Exclude variants that failed in Variant Quality Score Recalibration (VQSR), missing rate >2%, less than 80% of samples with depth <10x, the mean sequencing depth was <10 or >500, or the proportion of heterozygous genotypes among non-missing haploidy individuals >10%. (3) Exclude samples

with missing rate >5%, Het/Hom ratio >5, Ti/Tv ratio <2.5, the mean read depth <20x, the mean genotype quality is <55, or conflicted genetic and self-reported sex. Cryptic relatedness among samples were identified by the program KING. The individuals with the lowest missing rate among each group of duplicated or related samples were retained for subsequent analysis. We performed principal component analysis (PCA) to control for population stratification. We assessed whether the cases and controls were comparable by testing the differences in the number of rare variants per sample.

We annotated BayesDel score as a measure of the deleteriousness of each variant in the association test. We annotated functional consequences of variants against the RefSeq database, restricting to principal transcripts obtained from the APPRIS (annotation of principal and alternative splice isoforms) database. Loss-of-function (LoF) variants included FrameShifting, StopGain, and SplitSite variants. Any variants located in the last 5% of a coding region were not considered as LoF. Variants located within two base pairs in an intron were considered as SplitSite variants. For variants at -3 to +8 base pairs around a splicing donor site or at -12 to +2 base pairs around a splicing acceptor site, the splicing alternating effect was obtained from the dbscSNV database with the cutoff value of 0.9. We also conducted clinical annotation using the ClinVar database and excluded variants classified as benign. We conducted a gene-based burden test using a one-sided Fisher's Exact test within each population, restricting to rare variants with minor allele frequency or MAF <=0.01), and then a combined p-value was calculated by Fisher's method weighted by sample size.

Next, we developed a targeted sequencing panel for HNC using Phase 1 data. Using the VICTOR program, we perform three analyses on the first stage of data: (1) gene-based association test; (2) gene prioritization combining gene-based burden test results and biological relevant scores; and (3) gene set-based association test. We produced a prioritized gene list from each analysis. Using the VAAST and PheVor programs, we conducted another set of similar analyses, resulting in another three sets of prioritized gene lists.

Targeted gene sequencing of Phase 2 samples

Utilizing the panel of NHC risk genes designed using Phase 1 data, we performed targeted gene sequencing for all Phase 2 samples recruited from the three centers. The experiment was also conducted by the HCI High-Throughput Genomics Core Facility at the University of Utah Health Sciences. After obtaining the sequencing reads, we first trimmed FASTQ files using Cutadapt (Martin 2011) to remove residual adapters in the tails. We used FLASH (Magoc and Salzberg 2011) to combine the overlapped tail of the paired-end reads into a single long read if possible. Uncombined reads were treated as two single-end reads in the following analysis. The GATK Best Practices workflow (DePristo, Banks, et al. 2011) was applied for reference genome alignment, joint variant genotype calling, and variant quality score recalibration. The sequencing reads were mapped to the human reference genome (University of California, Santa Cruz or UCSC HG19) using Burrows-Wheeler Aligner (BWA) (v0.7.9a) (Li and Durbin 2009). We

removed unmapped reads, reads with mapping quality equal to 0, and duplicated reads using Picard tools (http://broadinstitute.github.io/picard). We conducted local realignment and minimized mismatched bases across reads using GATK's IndelRealigner. We recalibrated the base quality score and generated the final clean BAM file using GATK's BaseRecalibrator. We conducted joint variant genotype calling of all samples using GATK's HaplotypeCaller and variant quality score recalibration (VQSR) using GATK's VariantRecalibrator. To ensure the quality of the genotypes, we compared the results from the above two pipelines and only included the consistent genotype in the following analysis. For UK BioBank samples, we downloaded the individual gVCF files generated from whole exome sequencing (WES) from UK BioBank database. We utilized DeepVariant, which was implemented in GLnexus, to perform joint genotype calling.

After joint genotype calling, we conducted sample level and variant level quality control (QC) using the Cross-Platform Association Toolkit (XPAT)(Yu, Hu et al. 2018) for each dataset separately.

For sample-level quality control (QC), we computed the ratio of heterozygous variant counts to homozygous variant counts on autosomal chromosomes for each sample. Samples with ratios exceeding 3 were excluded. Additionally, we calculated NC90 scores for each sample using XPAT. NC90 scores represent the proportion of missing genotypes in a sample among all variants with call rates of 90% or greater across all samples, limited to variants within the target regions. Samples with NC90 scores greater than 0.05 were excluded. We also removed samples identified as 1st and 2nd-degree relatives predicted by RELPAIR (Epstein, Duren et al. 2000). Furthermore, we conducted Principal Component Analysis (PCA) to project all cases and controls onto a reference panel constructed using 1000 Genomes Phase 3 data (Genomes Project, Auton et al. 2015). This step was performed to identify genetically matched cases and controls for the following association test. For the target sequencing data of samples from three centers, we grouped the samples into four ethnicity groups: non-Hispanic European, African American, Asian, and Hispanic. For UK Biobank data, we only select cases and controls of non-Hispanic European ancestry.

We performed variant-level quality control (QC) for each population group of Utah samples and UK Biobank data, using XPAT with default parameters unless otherwise specified. Variants with VQSR tranche sensitivity scores >99.95 for SNPs and >99.5 for INDELs were excluded. Additionally, variants with an average allele fraction lower than 20% in our data were excluded. The average allele fraction was calculated as the proportion of reads supporting the minor allele among all reads in each sample with the minor allele. Variants with a missing genotype rate exceeding 10% across all samples were also excluded. Furthermore, variants with a p-value < 10^-6 in a Hardy-Weinberg equilibrium test in controls were excluded from further analysis. We conducted an internal PCA using cases and controls of each population and dataset separately, using well-behavior markers from the corresponding population (MAF > 20%), to capture

population stratification within each group. The first 10 components will be used as covariant in the following association tests.

Gene-based association analysis

For each of the four populations of Utah samples and the non-Hispanic European group of UK Biobank samples, we conducted the gene-based association test using three approaches. We first conducted gene-based association tests with all rare variants (minor allele frequency or MAF 0.005) using the Variant Annotation, Analysis and Search Tool (VAAST 2)(Yandell, Huff, et al. 2011, Hu, Huff et al. 2013). The analysis considered all coding variants, including missense, stop gains, stop losses, splicing variants, INDELs, and CNVs, weighting each variant according to the estimated degree of protein dysfunction conferred based on the CASM score in VAAST. We only test genes with at least four scorable variants. We then conducted the gene-based association with loss-of-function variants only, using a logistic regression method. The definition of excluded LOF variant. We also conducted an association test with both loss-of-function and predicted damaging variants. The definition of predicting damaging variants. For all gene-based tests, we used the first ten PCs, age, and sex as covariates to control for population and technological stratification. For each population group, we conducted an aggregated Cauchy association test (ACAT; Liu 2019) to combine the p values from individual tests. We next conducted a meta-analysis to combine the ACAT p value of each population group, using Stouffer's method. We used efficient sample size from each group as weight in the meta-analysis, defined as 4*number of cases*number of controls divided by the sum of cases and controls. We also conducted pathway-based analysis by combining p values of genes in each candidate pathway using Fisher’s method.

Effect size estimation for genes and pathways

We evaluated variant effect sizes of top genes in association tests for head and neck cancer risk according to a variety of functional classifications, considering their amino acid changes, *in silico* functional predictions, and variant annotation status in ClinVar using ANNOVAR (Version 20221231). For ClinVar annotations, we included ‘pathogenic’ and ‘likely pathogenic’ variants in the known pathogenic category, and uncertain or unknown significance as VUS. The effect size was estimated using the Firth regression method for each population group. We used the first ten PCs, age, and sex as covariates. To avoid bias due to the small sample size, we required at least 3 carriers of variants in cases and controls to estimate the odds ratio. For the non-Hispanic European group, we conducted the meta-analysis of the odds ratio to combine the estimates from Utah samples and UK Biobank samples {Med Care. 2017 Apr; 55(4): 328–335.}

Interaction analysis (Gene x Smoking, Gene x Drinking)

To assess epistasis between two predictors, we coded the two predictors into a four-level variable, corresponding to (1) negative for both predictors, (2) positive for the first predictor, (3) positive for the second predictor, and (4) positive for both predictors, respectively. Logistic regression was performed on this variable, using level 1 as the reference, adjusting for ancestry

and sex. Additive interaction between the two predictors was then assessed through the Synergy Index.

Supplementary Material

Experimental details of Phase 1 sequencing

Genomic DNA (gDNA) from Phase 1 samples were extracted from either peripheral blood samples or grossly uninvolved, normal formalin-fixed paraffin-embedded (FFPE) tissues. Libraries for DNA sequencing were prepared from 33ng of gDNA and were fragmented using a single-channeled Covaris S2 instrument to a target median fragment size of 250-300bps. KAPA Biosystem's HyperPrep library preparation system was used according to manufacturer's specifications, and samples were barcoded with adapters supplied by Roche-Nimblegen. DNA samples derived from FFPE tissues were treated with Uracil-DNA Glycosylase (UDG, New England Biolabs) by adding 0.005 units (1 uL) to the finished adapter-ligation reactions and incubating at 37°C for 15 minutes prior to the post-ligation purification. Barcoded libraries were multiplexed into sets of four and then target enrichment accomplished by hybridizing the libraries with Roche-Nimblegen's SeqCap EZ Human Exome Library v3.0 according to the manufacturer's protocol. Sequencing was conducted on sets of three multiplexed captures per channel on an Illumina HiSeq 2500 (125 cycle paired-end v4) by the HCI High-Throughput Genomics.

Sequencing results show a median read depth less than 100 on average (for blood-derived samples) or 150 (for FFPE samples) over the capture intervals specified by SeqCap EZ Human Exome v3.0 as determined by BEDTools (v2.22.1, github.com/arq5x/bedtools [4]) were requeued for additional sequencing to allow for the possibility of adequate coverage in aggregate. The sequence variant discovery pipeline was derived, with minor modifications, from the Broad Institute's GATK v3.3 best practices specification [5]. Recalibrating base quality scores, and generating single sample gVCF files using the HaplotypeCaller tool of GATK [6] with the following non-default parameters: emitRefConfidence = GVCF, stand_call_conf = 30.0, and min_base_quality_score = 20. Individual gVCF files from all of the sequenced individuals were aggregated along with suitable background gVCF (see below) using the CombineGVCF tool of GATK. Joint genotyping is performed using GATK's GenotypeGVCFs tools over the targeted capture regions. Variant scores were recalibrated using GATK VariantRecalibration with gold standard hapmap and 1000 genomes high confidence variants, and a 99.0 tranche threshold was applied to PASS or fail the genotype calls.