TGFβ receptor 1: an immune susceptibility gene in HPV-associated cancer

RUNNING TITLE: Integrative analysis of susceptibility to HPV-related cancer

Chaya Levovitz1,4,9, Dan Chen6, Emma Ivansson6, Ulf Gyllensten6, John Finnigan1, Sara Alshawish1, Wei-Ja Zhang7, Eric E. Schadt8, Marshal R. Posner1,2, Eric M. Genden1,2,5, Paolo Boffetta1,5,9, Andrew G. Sikora1,5

The Icahn School of Medicine at Mount Sinai1, Departments of Otolaryngology- Head and Neck Surgery2, Oncological Sciences3, Immunology4, Tisch Cancer Institute5, Department of Immunology, Genetics and Pathology and SciLifeLab Uppsala, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden6, Mount Sinai Department of Medicine7, Mount Sinai Institute for Genomics and Multiscale Biology8, Institute for Translational Epidemiology9

KEYWORDS: human papillomavirus, oropharyngeal cancer, integrative analysis, GWAS, TGFBR1

FINANCIAL SUPPORT: This work was partially supported by grants from the National Institutes of Health (NIH) National Cancer Institute (1F30CA165615-01) (CL); National Institute of Dental and Craniofacial Research (5R03DE021741-02) (AGS); the NIH National Cancer Institute (1K08CA154963-01A1) (AGS).

CONFLICT OF INTEREST: No potential conflicts of interest were disclosed

WORD COUNT: 5226
FIGURE COUNT: 4
TABLE COUNT: 3
SUPPLEMENTARY FIGURES: 3
SUPPLEMENTARY TABLES: 4

Coresponding Author:
Andrew G. Sikora. M.D., Ph.D.
Assistant Professor and Director of Head and Neck Translational Research
Department of Otolaryngology, Immunology, Oncological Science, and Dermatology
Mount Sinai School of Medicine
One Gustave L. Levy Place
Box 1189
andrew.sikora@mssm.edu (email)
212 659-9516 (Academic Office)
212 369-5701 (Fax)

ABSTRACT

[ABSTRACT Body]
Only a minority of those exposed to human papillomavirus (HPV) develop HPV-related cervical (CC) and oropharyngeal cancer (OPC). Since host immunity affects infection and progression to cancer, we tested the hypothesis that genetic variation in immune-related genes is a determinant of susceptibility to OPC and other HPV-associated cancers by performing a multitier integrative computational analysis with OPC data from a head and neck cancer (HNC) genome-wide association study (GWAS). Independent analyses, including single-gene, gene-interconnectivity, protein-protein interaction, gene expression, and pathway analysis, identified immune genes and pathways significantly associated with OPC. TGFBR1, which intersected all tiers of analysis and thus selected for validation, replicated significantly in the HNC GWAS limited to HPV-seropositive cases, and an independent CC GWAS. The TGFBR1 containing p38-MAPK pathway was significantly associated with OPC and CC, and TGFBR1 was overexpressed in OPC, CC, and HPV+HNC tumors. These concordant analyses implicate TGFBR1 signaling as a process dysregulated across HPV-related cancers. This study demonstrates that genetic variation in immune-related genes is associated with susceptibility to OPC, and implicates TGFBR1/TGFB signaling in the development of both OPC and CC. Better understanding of the immunogenetic basis of susceptibility to HPV-associated cancers may provide insight into host/virus interactions and immune processes dysregulated in the minority of HPV-exposed individuals who progress to cancer.

INTRODUCTION
Human papillomavirus (HPV) is a necessary cause of cervical cancer (CC), and a major cause of anal, vulvar, and penile cancer as well as oropharyngeal cancer (OPC) (1),(2),(3). While tobacco-associated squamous cell carcinomas of the head and neck (HNSCC) have declined in the United States (US) and other Western countries, the incidence of HPV-positive OPC has sharply increased in the US, rising by 225% since 1988, with over 70% of all newly-diagnosed OPC believed to be HPV positive (4),(5). The National Cancer Institute predicts that HPV-positive OPC will likely surpass CC as the most common HPV-associated cancer in the US by 2020 (5).

Although HPV infection is common (6), (7), only a small fraction of infected individuals develop cancer (8). The pattern of increasing relative risk with increasing degree of relatedness found in CC suggests that susceptibility to HPV-induced cancer is modulated by genetic factors (9). At the same time, the viral etiology of HPV-induced cancers implicates host immunity as a potential susceptibility factor (10), and the higher prevalence of cervical HPV infections, HPV positive CC (11) and HPV positive OPC (12) in HIV/AIDS individuals suggests that the host immune response is an important determinant of HPV-induced cancer risk. Thus we hypothesize that genetic variation in immune-related genes is a determinant of susceptibility to OPC and other HPV-associated cancers.

In the present study we investigate the immunogenetics of susceptibility to HPV-associated head and neck cancer using a genome-wide association study (GWAS) of upper aerodigestive (UADT) cancers that was organized by the International Agency for Research on Cancer (IARC) and the Centre National de Genotypage (CNG) in Paris and whose replication was conducted through the International Head and Neck Cancer Epidemiology (INHANCE) consortium (13). While the conventional unbiased GWAS approach can identify genetic variants associated with complex diseases (14), (15), many risk-modifying alleles are missed by this strategy due to the very stringent significance criteria required to minimize false discovery (16), (17). Alternate analytic approaches focused on combined effects of many loci, each contributing a small effect to the overall disease susceptibility, can successfully identify signals which individually do not meet the threshold for genome-wide significance but are in fact associated with the disease (18), (19). To test the hypothesis that variants in immune-related genes modulate susceptibility to OPC, an HPV-associated head and neck cancer, we developed a multitier integrative computational approach incorporating 4 distinct modes of analysis including: individual single-nucleotide polymorphism (SNP)/gene, pathway, gene-gene interconnectivity (GGI), and protein-protein interaction (PPI) analyses. Resulting hits that were supported across all four dimensions of analysis were subsequently validated in replication cohorts and gene expression studies (Figure 1). Our results are consistent with an important role for genetic variation in multiple immune-related genes as modifiers of HPV-related cancer risk, particularly those related to transforming growth factor beta (TGFB) signaling.

MATERIAL AND METHODS
IARC head and neck cancer GWAS and quality control of GWAS data

IARC conducted a large two-phased pooled case/control GWAS including 2091 UADT cancer patients (primarily head and neck cancer with esophageal cancers included) and 8334 cancer-free controls genotyped using the Illumina Sentrix HumanHap300 BeadChip. We requested the raw GWAS data from IARC, and subjected this data to additional cleaning and quality control (QC) before analysis. Details of the UADT GWAS including the study population and initial QC are reported in detail elsewhere (13). In brief, for this study, systematic QC was performed on the raw Illumina HumanHap300 genotyping data. We excluded samples with gender discrepancies, low call rates (<95%), and outliers in population structure from principle component analyses. Variants with a genotype call rate <95%, minor allele frequency (MAF) <5%, and deviation from Hardy-Weinberg equilibrium (P<10^-7) were removed. The extremely limited extent of admixture and population stratification was confirmed using multidimensional scaling analysis. The association between each genetic variant and the disease risk was estimated by the odds ratio (OR) per allele and 95% confidence interval (CI) using multivariate unconditional logistic regression assuming a log-additive genetic model with sex and country of recruitment included as covariates. All analyses were performed using PLINK (20). We conducted separate analyses on data restricted to OPC and laryngeal cancer (LC) subsites only.

Immune-related Gene Ranking and Significance Threshold

SNP identifiers were mapped to associated genes using the SNP location from the human genome assembly hg18. Gene boundary extensions, extending the transcription start and stop codon by 110kb and 40kb respectively, did not significantly change the results so they were not used in the final gene ranking in order to minimize error. The intersection between genes annotated from the set of SNPs and the 1324 genes defined by Gene Ontology (GO) network classification (21) as immune-related genes was found. Ranking was based on the most significant \( p \)-value for each SNP. Since immune-related genes are not independent of one another and operate in common pathways and networks, in our gene-based analysis we used an false discovery rate (FDR) <0.05 for a significance threshold.

In-silico replication of OPC GWAS results in serologically HPV positive head and neck cancer patients

A Luminex-based multiplex immunosorbent method was used to determine each subject’s antibody response to HPV infection by measuring serologic reactivity to the HPV16 E6 protein, as previously described (22). The cutoff was defined as 5 standard deviations above the mean of the final distribution. The serology data was then used to perform a validation analysis on 131 serologically HPV positive head and neck cancer patients and 2919 HPV negative controls using the same methodologies applied in the OPC GWAS (described above). A significance cutoff of \( p < 0.05 \) was used in the validation.

Gene-Gene Interconnectivity analysis

Downloaded from cancerres.aacrjournals.org on April 13, 2017. © 2014 American Association for Cancer Research.
The highly ranked immune-related genes, corresponding to SNPs with p-values suggestive of association (1x10^{-3} > P >1x10^{-7}) were analyzed in Gene Relationships Across Implicated Loci, GRAIL (23), a literature-based text-mining program that infers biological interconnectivity between and among genes. In summary, genes were identified using the linkage disequilibrium (LD) structure of HapMap2 Release 22 CEU and text mining was performed using a text-based similarity measure that scores two genes for relatedness to each other based on text similarity in PubMed abstracts last curated in 2012. Genes in regions of interest were clustered based on keyword similarity. These clusters were then scored based on ranked similarity, adjusting for gene size, to generate p-values evaluating the strength of the functional interconnectivity of genes in the regions of interest. A significant p-value for a gene region indicates that a gene within it is more related to genes in other disease regions through PubMed abstracts than expected by chance. P-values for these functional clusters were FDR adjusted to correct for multiple testing with the FDR<0.05 considered the threshold of significance.

**Protein-Protein-Interaction analysis**

Protein-protein interaction analysis was completed using Disease Association Protein-Protein Link Evaluator, DAPPLE. The details of the algorithm and methods used in DAPPLE can be found in Rossin EJ et al (24). Briefly, the top 200 ranked immune-related SNPs (maximum number allowed by the program) were input into DAPPLE as seed SNPs and converted into genes. DAPPLE then built interaction networks from proteins encoded by the seed genes that have direct connections reported in the literature. The statistical significance was assessed by a number of network connectivity parameters as well as of the connectivity of the individual proteins to other seed proteins using a within-degree node-label permutation method. The individual p-values for seed proteins represent the probability that by chance the seed protein would be as connected to other seed proteins at the level observed in the network. Clusters were assigned based on statistically significant connections of genes that participate in common biological functions.

**Pathway analysis**

We identified pathway-based associations in the OPC GWAS with the Meta-Analysis Gene-set Enrichment of variant Associations (MAGENTA) platform (25). The gene set enrichment (GSEA)-based methodology is described in Segre’ et al. In summary, all SNP from the GWAS analysis limited to OPC were used in the pathway analysis. Genes in the human genome were mapped to a single index SNP with the lowest p-value. This p-value became the gene score and was corrected for confounding factors such as gene size, SNP density and LD-related properties in a regression model to determine gene-wise adjusted gene score. Genes were then ranked by the adjusted gene scores. At a significance threshold of 95th and 75th percentile of all gene scores, the observed number of gene scores in a given predefined biological pathway, with a ranked score above the specified threshold percentile, was calculated. This observed statistic was then compared to 1,000,000 randomly permuted pathways of identical size and generates an empirical GSEA p-value for each pathway. Significance
was determined when a pathway reached a FDR <0.05 for either threshold. Predefined biological pathways were taken from several publically available databases including GO(N=9431), PANTHER(N=634), KEGG(N=185), Ingenuity(N=92), BIOCARTA(N=217), REACTOME(N=427), and WikiPathways(N=43).

**In-silico replication of OPC results in a CC GWAS**

An *in-silico* replication was performed using a GWAS of 617 unrelated CC patients and 512 cancer-free controls generated on the Affymetrix Genome-wide Human SNP array 5.0 with ~440,794 SNP markers, which uses mostly different SNP markers compared to the Illumina arrays used in the UADT cancer GWAS. The details of population and QC have been described in Ivansson EL et al. (26). The data we received included for each SNP, the SNP identifier, the genomic location, and the p-value from the association analysis that was estimated by ORs and 95% CI using multivariable unconditional logistic regression assuming a log-additive genetic model. Single gene analysis was completed as described above. A significance cutoff of $p<0.05$ was used in the validation. A combined p-value was calculated for each significant gene using Fishers method. Using MAGENTA, a pathway analysis limited to pathways found to be significant in the OPC dataset was also performed on the CC GWAS dataset.

**Gene expression analysis**

Oncomine was used for analysis and visualization of tumor gene expression data (27). Oncomine is an online tool that aggregates mRNA expression data from a large number of cancer gene expression microarrays, and allows the user to investigate the relative expression of genes across various datasets. The relative expression of significant genes was investigated across OPC and CC datasets. The overall $p$-values were determined by simultaneously considering available data within Oncomine for the cancer versus normal comparisons. $p$-values $<0.05$ was considered significant. Only studies based on human samples were included in the analysis.

A parallel analysis was performed using normalized level 3 gene expression data for all HNSCC samples extracted from The Cancer Genome Atlas (TCGA) RNASeqV2 protocol (28, 29). 69 HPV-positive cases were identified, with HPV status determined by cross-referencing the corresponding TCGA clinical dataset for samples whose HPV status had been verified by in situ hybridization or PCR testing. 42 unmatched samples corresponding to surrounding non-neoplastic tissue were also identified. A 2 tailed t-test was used to assess statistically significant differential expression of TGFBR1 for all samples. P-value $< 0.05$ was considered significant.

**RESULTS**

**Single-SNP and Single-gene Association Analysis**
After stringent quality control, data from 317 OPC cases and 3707 controls were tested for association with 296728 SNPs. The QQ plot (Supplementary Figure S1) shows minimal evidence of genomic inflation \((\lambda = 1.01539)\). To enhance the power to detect associations, we dramatically reduced the number of SNPs tested by limiting analysis to the 12258 SNPs that map to Gene Ontology (GO) (21) categorized immune-related genes in order to test the hypothesis that germline DNA variants in these genes modulate risk of OPC. Of the 1324 GO-annotated immune-related genes, 1304 could be assigned to a SNP. Of the top ranked immune-genes corresponding to SNPs with \(p\)-values suggestive of association \((1 \times 10^{-3} > p > 1 \times 10^{-7})\), the topmost 10 were significant at a gene-based level after correction for multiple testing \((\text{FDR} < 0.05)\), while all have been previously associated in the literature with cancer, viral infection, or both (Table 1).

To confirm the oropharyngeal subsite is a valid surrogate for HPV-related head and neck cancer, we performed a validation analysis on serologically HPV positive head and neck cancer patients. As expected, HPV E6 seropositivity was predominantly found in patients with OPC (Supplementary Table S1, Supplementary Figure S2). Even with the statistical significance limited by the low numbers of HPV positive cases and limited sensitivity of serological testing (30),(31) the overall results were strikingly supportive of the association analysis restricted to OPC. As shown in Table 1 and Supplementary Table S2, all 14 immune-related genes identified in the OPC GWAS with SNP \(p\)-values suggestive of association replicated significantly in the HPV+ cohort \((p\text{-value} < 0.05)\).

To determine whether immune-related genes are uniquely associated with OPC rather than all types of HNSCC, we performed a parallel analysis of laryngeal cancer (LC), a head and neck cancer strongly associated with tobacco and alcohol use rather than HPV infection (32). As shown in Table 1, out of the 14 highly ranked genes in the OPC analysis that replicated in the HPV+ cohort, only 1 replicated in LC. The Two-sample Kolmogorov-Smirnov test demonstrated that the distribution of \(p\)-values of the candidate immune-related genes was much more significant \((p\text{-value} = 6.64E-04)\) in the OPC analysis compared with the LC analysis, demonstrating a unique association of candidate immune genes with OPC that is not seen with tobacco/alcohol-associated head and neck cancer. Association of candidate genes was also much higher for OPC than for all HNSCC aggregated (data not shown).

In summary: the suggestive association of OPC with multiple candidate genes plausibly linked to viral immunity, the robustness of these associations when applied to HPV seropositive head and neck cancer patients, and the lack of similar strong association of immune related genes with non-HPV-related head and neck cancer supports our hypothesis that there is an immunogenetic basis for OPC risk due to its unique association with HPV. We then further tested this hypothesis by applying further tiers of validation to the genes most strongly associated with OPC risk.

**Gene-gene Interconnectivity and Protein-protein interaction (PPI) analysis**
Analysis of functionally-related gene combinations can unmask cumulative/cooperative risk associations that would be missed by single-gene analysis. To explore functional and biological processes driving susceptibility to OPC we performed gene-gene interconnectivity and protein-protein interaction analyses on immune-related genes from the OPC single-gene association analysis. Gene-gene interconnectivity analysis demonstrated a high degree of significant interconnectedness among these genes, suggesting their involvement in common biological processes (Figure 2A and Supplementary Table S3). PPI analysis performed on the top 200 ranked SNPs in immune-related genes from the OPC analysis revealed multiple statistically significant first-degree interactions (Figure 2B and Supplementary Table S4). The PPI network contained 81 direct connections between seed proteins from different loci with an expected value of only 24.58 ($p$-value = 9.99E-04) and the average seed protein direct binding degree of 2.7 with an expected value of only 1.42 ($p$-value = 9.99E-04). In randomly selected subsets of 200 genes from the set of 1304 GO-annotated immune-related genes, the connectivity in the PPI network was significantly reduced and both the direct connections between seed proteins and the average seed protein direct binding degree were not significant. (Supplementary Figure S3A).

Both the GGI and PPI analyses indicate that the top-ranked immune-related genes display a high degree of interconnectivity and are likely to be involved in related biological processes. Visual inspection of the PPI plot identified 3 distinct clusters related to TGFB signaling, Th1/Th2 balance, and innate immunity. Even when we widened the gene margins to be less stringent by extending the transcription start and stop codon by 110kb and 40kb respectively (Supplementary Figure S3B), these clusters are apparent, demonstrating their robustness.

**Pathway analysis**

To identify specific immune pathways that may play a role in the etiology of HPV-mediated cancer we performed a gene set enrichment analysis (GSEA) of 11,029 pathways associated with a broad array of biological processes. We identified 21 immune-related pathways significantly associated with OPC, organized into 6 categories including 5 toll-like receptor (TLR) and innate immunity pathways, 4 NFKB-related pathways, 2 T cell activation pathways, 4 cell death pathways, 4 TGFB/TGFBR-related pathways, and 2 miscellaneous immune pathways (Table 2). Many of these pathways contained one or more of the top ranked genes from the OPC single-gene association analysis, most notably the p38 MAPK pathway which contains MAPKAPK2 and for which TGFBR1 is the most significant gene. Most strikingly, many of the highly ranked immune-related genes found in the single-gene analysis participated in significant PPI and/or significant pathways including TGFBR1, IL10, SOCS5, ZAP70, MAPKAPK2, CD80, LYN and MAPK10. Several immune-related genes that displayed significant PPI were also the most significant genes in many of the significant pathways from GSEA analysis, including TGFBR1, EGFR, CD44, IGF1R and JAK2 (Figure 3). The density of participation of highly-ranked immune related genes in functional pathways and protein-protein and gene-gene interaction networks is consistent with our hypothesis that multiple functionally-related genes participating in key immune processes contribute incrementally to risk of OPC.

**Replication of OPC results in a CC GWAS**

Downloaded from cancerres.aacrjournals.org on April 13, 2017. © 2014 American Association for Cancer Research.
Our hypothesis that risk of OPC, as an HPV-related cancer, is uniquely associated with variation in immune genes, would be strongly supported by validation in CC, the cancer most strongly associated with HPV. Since TGFBR1 was significant across all 4 tiers of independent analyses, we attempted to replicate it in the CC GWAS. First, we sought to confirm that variants in TGFBR1 are associated with CC, and found that the TGFBR1-associated SNP rs334356 was ranked the 11th most significant immune gene in the CC GWAS with a p-value=2.00E-03. Since a SNP-based replication was unable to be performed because the OPC and CC GWAS each used different genotyping platforms, we attempted a gene-based replication. The combined p-value of TGFBR1 across the OPC and CC GWAS met the significance threshold for gene-based analysis (using the nearest gene and most significant SNP in each gene loci) even after standard Bonferroni correction (p<3.8E-05) with a p-combined = 1.25E-05.

We also sought to determine whether any of the significant pathways from the OPC analysis replicated in the CC dataset. The P38 MAPK pathway was the only pathway significant in both the OPC and CC GWAS datasets (FDR adjusted p-value <0.05), and in both datasets TGFBR1 was the most significant gene. Of the 54 genes comprising the p38 MAPK pathway, the same seven genes had the lowest p-values and were drivers of significance in both the OPC and CC GWAS datasets (Figure 4A), suggesting conservation of functional gene associations across both HPV related cancers. To determine whether TGFBR1-related pathways are uniquely associated with HPV-related cancers we compared the nominal p-values and corresponding FDR of all immune-related pathways containing closely TGFBR1-related genes (TGFB1, TGFB2, TGFB3, TGFBR1, TGFBR2, TGFBR3) in OPC and CC (HPV related) and LC (tobacco/alcohol related; Figure 4B). The significance trends are more similar for the two HPV-mediated cancers, OPC and CC, than either is to LC, which shows almost inverse significance trends. These findings support a unique and potentially functionally-conserved role for TGFB signaling across HPV-related cancers.

**Expression of TGFBR1 is altered in HPV-associated tumors**

Significant genetic associations with OPC and other HPV-associated cancers are likely to have functional consequences, such as altered expression of target genes in cancer. We analyzed TGFBR1 gene expression in OPC and CC tumor specimens in the Oncomine human cancer genomic database (27) and in HPV+ head and neck cancer samples in the TCGA database (28, 29). TGFBR1 was significantly overexpressed in OPC, CC, and HPV+ head and neck cancer with respect to benign tissue (Table 3). This suggests that the genes found to be associated with HPV-related head and neck cancer are not simply markers of susceptibility, but likely to play a functional role in interactions between the host and the virus or virally-transformed cancer cells.

**DISCUSSION**
Little is known about determinants of host susceptibility to viral carcinogenesis, and whether common biological themes may account for the heterogeneity of progression to cancer following viral exposure. The present study is the first to exploit GWAS as a high-throughput strategy to examine immunogenetic susceptibility to OPC, unique among HNSCC subsites for its strong association with HPV. We found that variation in immune-related genes is an important determinant of susceptibility to HPV-related but not HPV-unrelated HNSCC, and that this relationship is robust across multiple levels of analysis. Our findings also specifically highlight the pivotal contribution of variation in TGFB/TGFBR signaling-associated molecules to HPV-related cancer susceptibility, which was consistent across OPC, HPV seropositive HNSCC and CC analyses. Significance of immune-related genes was much greater in the OPC GWAS than that in a parallel analysis of LC, consistent with a unique contribution of immune-related genes to HPV-related cancer. While this does not mean that host immunity plays no role in susceptibility to LC or other tobacco-associated HNSCC, it highlights the impressive and unique magnitude of association between variation in specific immune-related genes and OPC, presumably due to high proportion of virally-induced cancers at this subsite.

The strongest support for GWAS findings is their replication in independent datasets. For this we focused on TGFBR1 and the TGFB signaling pathway, which were significant across all modes of analysis. Since the association between immune genes and risk of OPC is presumed to reflect enrichment at this subsite for HPV-driven cancer, we replicated our findings in CC, another HPV-related cancer. TGFBR1 was successfully validated in the CC GWAS, with a significance level in the top 99.999% of the whole GWAS. Pathway level analysis comparing significance among OPC, CC, and LC of all pathways that include a core TGFB-related gene, also strongly supported a role for TGFBR1 and TGFB signaling unique to virally-associated cancer. The likely functional significance of altered TGFB signaling in HPV-related cancers is further supported by our finding that TGFBR1 is significantly overexpressed in both OPC and CC.

These findings are consistent with the existing literature, in which TGFB signaling has already been linked to HPV-associated cancer by classical methods of genetic analysis. Dysregulated TGFB signaling is associated with malignant progression of HPV positive cervical dysplasia (33), (34), and HPV has been shown to promote CC by attenuating TGFBR1 signaling required for epithelial homeostasis at early stages of viral infection (35). Most importantly, genetic variants in TGFBI have also been shown to be associated with HPV positive OPC, with people carrying genotypes with TGFBI variants more than twice as likely to have an HPV positive tumor as patients with the wild type genotype (36). Our results further support the concept that dysregulated TGFB signaling is a key process common to multiple HPV-related cancers.

In addition to TGFBR1, all the genes linked to the top-ranked SNPs are highly plausible candidates with established connection to cancer and/or viral infection (Figure 3 and Table 1), and several have been previously reported to be associated with CC including CRTAM, CD80, IL10 and EBF1 (37) (38),(39),(40). Over half of the top-ranked immune genes were significantly associated with OPC across multiple tiers of analysis,
supporting the hypothesis that the class of immune-related genes is uniquely enriched for OPC susceptibility genes. While these associations at the single-SNP level do not meet the criteria for genome-wide significance \((p<1\times10^{-8})\), the immune system is characterized by interconnectivity, cooperativity, and redundancy of functions. Thus even in the absence of individual SNPs meeting genome-wide significance, immune-related genes may contribute additively and incrementally to cancer risk by participating in common pathways and networks underlying disease susceptibility. Hence our decision not to focus on single SNPs, but to consider their significance in the context of a broad, biology-driven, pathway/network-discovery approach. The broad concordance of our results across disparate tiers of analysis including gene-gene interconnectivity, protein-protein interaction, GSEA/pathway, and gene expression analyses supports the premise that functionally-related gene sets rather than individual genes modulate susceptibility to OPC.

Additional evidence from the literature supports the association of multiple immune related genes with susceptibility to HPV-related cancer. In a recent GWAS of CC in a Swedish population (41) three independent loci were identified in the major histocompatibility complex (MHC) region that influence susceptibility, the first in the MHC class I polypeptide-related sequence A gene \((MICA)\), the second between \(HLA-DRB1\) and \(HLA-DQA1\), and the third at \(HLA-DPB2\). Another CC GWAS performed in China also found associations with the \(HLA-DPB1\) and \(HLA-DPB2\) genes (42). Further support for a role for HLA molecules as modulators of susceptibility to CC comes from a pathway analysis performed on the CC GWAS which a subset was used as replication cohort in this study and identified several pathways including HLA genes that influence risk of CC (26). However, in the OPC GWAS, the pathways and SNPs associated with human leukocyte antigen (HLA) molecules were not highly ranked, with the most significant SNP \(p\)-value equal to 1\(\times10^{-2}\). These differences may potentially reflect differences in the underlying biology of OPC and CC, which share HPV as an etiologic agent, but are in many respects clinically and epidemiologically different. On the other hand, a recent study (43) found NFkB-related pathways to be significantly associated with CC and vulvar cancer, which is concordant with our identification of 4 NFkB-related pathways in OPC.

One significant limitation of this study is the lack of gold-standard tumor HPV status, requiring us to use HPV serology to infer which patients have HPV-associated tumors. While HPV serology of blood samples is an established method for identifying patients with a high likelihood of HPV-induced cancer, the methods used for serologic analysis only identify roughly 60% of HPV+ OPC (31), (30). An additional limitation of the approach used is that we only considered seropositivity for HPV16, while other oncogenic subtypes are associated with OPC and presumably interact with the immune system in a similar fashion. Thus we based our primary analysis on OPC, the cancer subsite most highly associated with HPV, and used serology results to support the unique association of OPC with HPV-related cancer. While limited by the low number of cases with positive serology, the OPC and HPV seropositive datasets had strongly overlapping highly ranked significant SNPs and immune-related genes. Also, our results showing a much stronger association of immune genes with OPC as compared...
to LC and all HNSCC, and the close similarity to CC, further support our hypothesis that the association of variation in immune-related genes with OPC is uniquely strong because of its association with HPV-induced cancer. Replicating our OPC findings in the CC dataset further validates our findings as genetic signatures associated with HPV-driven carcinogenesis.

In summary, this paper supported the hypothesis that variants in immune-related genes, pathways and networks increase susceptibility to HPV-associated HNSCC. Using a multitier analytic approach, TGFB signaling was found to be associated with both HPV-mediated HNSCC and CC. Although this study focused on replicating the finding with the most analytical evidence, TGFB signaling, we plan to investigate other immune-related genes and pathways that showed striking overlap between tiers of analysis. Identifying these causal variants will provide clues to pathogenetic mechanisms against which preventative interventions might be targeted and novel immunotherapeutic strategies developed for HPV-mediated OPC.

ACKNOWLEDGMENTS:

We would like to thank Dr. Inga Peter (Department of Genetics and Genomic Sciences, ISMMS) and Dr. Ke Hao (Department of Genetics and Genomic Sciences, ISMMS) for their helpful input.

REFERENCES:


TABLES

Downloaded from cancerres.aacrjournals.org on April 13, 2017. © 2014 American Association for Cancer Research.
# Table 1: Immune genes with p-values suggestive of association

<table>
<thead>
<tr>
<th>SNP</th>
<th>GENE</th>
<th>OPC P value</th>
<th>OR</th>
<th>CI</th>
<th>Rank</th>
<th>HPV+ P value</th>
<th>Larynx P value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9922120</td>
<td>CBFA2T3</td>
<td>8.10E-07</td>
<td>1.646</td>
<td>1.35-2.01</td>
<td>6</td>
<td>4.52E-03</td>
<td>1.16E-02</td>
<td>Candidate breast cancer tumor suppressor(44) Associated with AML(45)</td>
</tr>
<tr>
<td>rs895437</td>
<td>ZAP70</td>
<td>6.64E-05</td>
<td>1.653</td>
<td>1.29-2.12</td>
<td>40</td>
<td>6.90E-03</td>
<td>7.25E-01</td>
<td>Mutations cause selective T-cell defect(48) Used as a prognostic marker for CLL in B cells(49) Mutations results in immune deficiency(50)</td>
</tr>
<tr>
<td>rs3107638</td>
<td>CRTAM</td>
<td>8.75E-05</td>
<td>1.551</td>
<td>1.25-1.93</td>
<td>51</td>
<td>5.15E-04</td>
<td>2.87E-01</td>
<td>KO mice have reduced viral immunity(51) Associated with esophageal SCCA(53) Induces apoptosis(54) Retards ovarian tumor growth(55)</td>
</tr>
<tr>
<td>rs1780365</td>
<td>PRK1</td>
<td>2.09E-04</td>
<td>0.711</td>
<td>0.59-0.85</td>
<td>113</td>
<td>8.28E-04</td>
<td>6.81E-01</td>
<td>Activated by pro-inflammatory cytokines(56) Associated with cancer and inflammation(57)</td>
</tr>
<tr>
<td>rs2869433</td>
<td>MAPK10</td>
<td>2.80E-04</td>
<td>0.708</td>
<td>0.59-0.85</td>
<td>134</td>
<td>4.66E-02</td>
<td>9.66E-01</td>
<td>Necessary for T cell activation and survival(58) Associated with cancer including cervical cancer(38)</td>
</tr>
<tr>
<td>rs6778945</td>
<td>CD80</td>
<td>3.75E-04</td>
<td>0.590</td>
<td>0.44-0.79</td>
<td>171</td>
<td>5.85E-03</td>
<td>8.01E-01</td>
<td>Involved in immunoregulation and inflammation(59) Released during immune response to viral infection(60) Associated with cervical cancer [PMID: 18341210]</td>
</tr>
<tr>
<td>rs3024498</td>
<td>IL10</td>
<td>3.95E-04</td>
<td>1.421</td>
<td>1.17-1.73</td>
<td>181</td>
<td>6.75E-04</td>
<td>3.03E-01</td>
<td>Involved in regulating T helper cell differentiation(61) Associated with HIV replication(62) Suppresses EGFR signaling (63)</td>
</tr>
<tr>
<td>rs10167561</td>
<td>SOCS5</td>
<td>4.23E-04</td>
<td>0.710</td>
<td>0.59-0.86</td>
<td>192</td>
<td>2.00E-02</td>
<td>3.32E-01</td>
<td>Pro-oncogenic role(64) Advanced carcinomas and tumor invasiveness(65) Polymorphism associated with several cancers(65)</td>
</tr>
<tr>
<td>rs2026811</td>
<td>TGFB1</td>
<td>4.29E-04</td>
<td>1.414</td>
<td>1.17-1.72</td>
<td>196</td>
<td>1.85E-03</td>
<td>5.23E-01</td>
<td>Associated with cervical cancer(40) Associated with HNSCCA (66) Associated with breast cancer(67)</td>
</tr>
<tr>
<td>rs11135045</td>
<td>EBF1</td>
<td>6.71E-04</td>
<td>1.437</td>
<td>1.16-1.77</td>
<td>289</td>
<td>1.68E-02</td>
<td>2.68E-01</td>
<td>Associated with cervical cancer(40) Associated with HNSCCA (66) Associated with breast cancer(67)</td>
</tr>
<tr>
<td>rs11633294</td>
<td>IGF1R</td>
<td>7.03E-04</td>
<td>0.725</td>
<td>0.60-0.87</td>
<td>310</td>
<td>4.09E-03</td>
<td>5.27E-01</td>
<td>Anti-apoptotic (68) Associated with cancer (68)</td>
</tr>
<tr>
<td>rs2667975</td>
<td>LYN</td>
<td>7.56E-04</td>
<td>0.706</td>
<td>0.58-0.86</td>
<td>331</td>
<td>1.57E-03</td>
<td>1.78E-01</td>
<td>Associated with myeloid progenitor and monocyte tumors(69) Associated with breast cancer(70)</td>
</tr>
<tr>
<td>rs3816375</td>
<td>ITGAV</td>
<td>7.75E-04</td>
<td>1.335</td>
<td>1.13-1.58</td>
<td>338</td>
<td>4.90E-02</td>
<td>3.11E-01</td>
<td>Associated with HCC(71); In HIV, associated with Kaposi sarcoma lesions(72)</td>
</tr>
</tbody>
</table>

# Table 2: Significant pathways
<table>
<thead>
<tr>
<th>Immune Category</th>
<th>Database</th>
<th>Pathway</th>
<th>FDR</th>
<th>Pathway GS</th>
<th>Effective GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell death</td>
<td>GOTERM</td>
<td>Negativa regulation of caspase activity</td>
<td>2.98E-02</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Cell death</td>
<td>PANTHER</td>
<td>Other apoptosis</td>
<td>1.54E-02</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Cell death</td>
<td>REACTOME</td>
<td>P75ntr signals via NFkB</td>
<td>2.01E-02</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Cell death</td>
<td>REACTOME</td>
<td>Nrif signals cell death from the nucleus</td>
<td>3.23E-02</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>MISC immune</td>
<td>PANTHER</td>
<td>B cell and antibody-mediated immunity</td>
<td>3.66E-02</td>
<td>97</td>
<td>65</td>
</tr>
<tr>
<td>MISC immune</td>
<td>GOTERM</td>
<td>initiation of viral infection</td>
<td>3.74E-02</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>NFKB</td>
<td>BIOCARTA</td>
<td>EPONFKB pathway</td>
<td>5.20E-03</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>NFKB</td>
<td>Ingenuity</td>
<td>NFKB signaling</td>
<td>3.05E-02</td>
<td>43</td>
<td>39</td>
</tr>
<tr>
<td>NFKB</td>
<td>REACTOME</td>
<td>Human TAK1 activates NFKB by phosphorylation and activation of IKKS complex</td>
<td>4.00E-02</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>NFKB</td>
<td>BIOCARTA</td>
<td>NFKB pathway</td>
<td>3.90E-03</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>T Cell Activation</td>
<td>BIOCARTA</td>
<td>CD40 pathway</td>
<td>2.21E-02</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>T Cell Activation</td>
<td>Panther</td>
<td>T cell activation</td>
<td>2.23E-02</td>
<td>31</td>
<td>22</td>
</tr>
<tr>
<td>TGFβ signaling</td>
<td>wiki pathway</td>
<td>P38 MAPK signaling pathway</td>
<td>4.56E-02</td>
<td>54</td>
<td>52</td>
</tr>
<tr>
<td>TGFβ signaling</td>
<td>KEGG</td>
<td>MAPK signaling pathway</td>
<td>2.19E-02</td>
<td>267</td>
<td>241</td>
</tr>
<tr>
<td>TGFβ signaling</td>
<td>wiki pathway</td>
<td>MAPK signaling pathway</td>
<td>4.80E-02</td>
<td>34</td>
<td>33</td>
</tr>
<tr>
<td>TGFβ signaling</td>
<td>wiki pathway</td>
<td>TGFβ signaling pathway</td>
<td>4.62E-02</td>
<td>162</td>
<td>154</td>
</tr>
<tr>
<td>TLR and Innate Immunity</td>
<td>REACTOME</td>
<td>Toll like receptor 3 cascade</td>
<td>1.61E-02</td>
<td>59</td>
<td>55</td>
</tr>
<tr>
<td>TLR and Innate Immunity</td>
<td>Ingenuity</td>
<td>LPSIL-1 mediated inhibition of RXR function</td>
<td>1.80E-02</td>
<td>59</td>
<td>55</td>
</tr>
<tr>
<td>TLR and Innate Immunity</td>
<td>KEGG</td>
<td>NOD like receptor signaling pathway</td>
<td>4.91E-02</td>
<td>62</td>
<td>52</td>
</tr>
<tr>
<td>TLR and Innate Immunity</td>
<td>Panther</td>
<td>Toll receptor signaling pathway</td>
<td>4.26E-02</td>
<td>30</td>
<td>26</td>
</tr>
<tr>
<td>TLR and Innate Immunity</td>
<td>REACTOME</td>
<td>Toll receptor cascades</td>
<td>2.03E-02</td>
<td>86</td>
<td>78</td>
</tr>
</tbody>
</table>
Table 3: **TGFB1 gene expression analysis**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>p-value</th>
<th>Fold Change</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oropharyngeal Cancer</td>
<td>2.00E-03</td>
<td>2.21</td>
<td>Oncomine</td>
</tr>
<tr>
<td>Cervical cancer</td>
<td>3.85E-07</td>
<td>2.88</td>
<td>Oncomine</td>
</tr>
<tr>
<td>HPV+ head and neck cancer</td>
<td>1.04E-08</td>
<td>1.69</td>
<td>TCGA</td>
</tr>
</tbody>
</table>
Table 1: **Immune genes with p-values suggestive of association.** The table lists the top 14 ranked immune genes and their corresponding SNP and includes the SNP identifier (rs#), the official gene symbol, the overall rank in the OPC GWAS, the genes most significant SNP’s p value from the OPC, HPV+, and LC association analysis and connection of each gene to cancer, progressive viral infection, or both described in existing scientific literature. The top 10 ranked genes are significant at a gene-based threshold (FDR<0.05).

Table 2: **Significant pathways.** The 21 significant immune pathways assigned to one of 6 immune categories. For each pathway the table lists the immune category, the database the pathway was found in, the official name of the pathway, FDR, and the gene size (GS, the number of genes included in the pathway) and the effective gene size (EGS, the number of genes found in the GWAS dataset).

Table 3: **TGFBRI gene expression analysis.** Gene expression levels of TGFBRI in OPC, CC, and HPV+ HNSCC. The table lists the tissue type, p-value, fold change, and database for all entries.

Figure 1: **Work flow for the multitier computational project.** Quality control and association analysis was performed on the GWAS raw data, which was then subjected to several independent modes of analysis. Validation was performed on genes and pathways found to be significant in all modes of analysis.

Figure 2: **Literature based connectivity analyses.** (A) GGI analysis demonstrates the high degree of significant connectivity amongst the highly ranked immune SNPs in the OPC GWAS. The figure was generated using the GRAIL platform, which determined the corresponding corrected p-values for all gene-gene connections. The outer ring contains the input SNPs and the inner ring shows all possible genes corresponding to each SNP locus. Genes with significant inter-gene connections, determined by the GRAIL p-value (Supplementary Table S2), are in bold and the thickness of the connection corresponds to strength of their inter-connections. (B) PPI network for the top ranked 200 immune genes in the OPC GWAS generated using the DAPPLE platform. Nodes correspond to proteins, links correspond to direct interactions reported in the literature and colors correspond to the significance of the interaction as defined by the color key. Visual analysis of the OPC GWAS immune network reveals 3 distinct immune clusters: (1) TGFB signaling (2) Innate immunity (3) Th1/Th2 balance.

Figure 3: **The intersections of the integrative analysis.** The results of each mode of analysis are listed in the Venn diagram. Multiple genes were found to overlap the results of 2 analyses but only TGFBRI intersected the results of all three independent analyses. (Single gene analysis was not included in this diagram as results informed the GGI analysis and thus would be redundant.)

Figure 4: **Pathway replication.** (A) The P38 MAPK pathway containing TGFBRI that is significant in both the OPC and CC GWAS. Significant p-values from the OPC and CC datasets are listed for each gene. The colors correspond to the dataset in which the genes were significant, as described in the color key. TGFBRI is the most significant gene in both datasets. (B) The comparative significance of all TGFB related pathways in the 2 HPV-associated cancers, OPC and CC, and in classical non-HPV mediated HNSCC, laryngeal cancer. Heatmap color intensity is proportional to the p value or FDR as indicated. Pathway names highlighted in red represent those that include TGFBRI.

**SUPPLEMENTARY FIGURE AND TABLE LEGENDS:**
Supplementary Table S1: **Normalized Average E6 Seroreactivity in HNC Subsites**

Supplementary Table S2: **Immune genes in the HPV+ analysis.** The table lists the most significant SNPs and corresponding immune genes from the HPV+ analysis. Data includes the immune gene name, SNP identifier, odd ratio (OR), confidence interval (CI), and p-value (p)

Supplementary Table S3: **GGI analysis p-values determined using GRAIL.**

Supplementary Table S4: **PPI analysis p-values determined using DAPPLE.**

Supplementary Figure S1: **OPC GWAS Analysis** (A) Manhattan plot of OPC GWAS analysis. (B) QQ plot for OPC GWAS analysis.

Supplementary Figure S2: **Comparative scatter plot by HNSCC tumor site vs HPV E6 serology.**

Supplementary Figure S3: **DAPPLE randomization and extended boundary analysis.** (A) A random set of 200 immune genes, using the wide margin (less stringent) algorithm for SNP-gene assignment, does not demonstrate the same clustering effect in PPI analysis as does the set of highly-ranked immune genes in the OPC GWAS, confirming that the ranking order is specific and functionally/biologically relevant to OPC. (B) While transcription start and stop margins of genes (as reported in hg18) are extended by 110kb and 40kb respectively in the OPC dataset, the three clusters are still evident.
Figure 1: Work flow for the multitier computational project.
Figure 2: Literature based connectivity analyses
Figure 3: The intersections of the integrative analysis.
Figure 4: Pathway replication.

A

B

OPC and CC Pathway Driver
Significant in OPC and CC
Significant only in OPC
Not significant in OPC or CC

negative regulation of apoptosis 1
negative regulation of apoptosis 2
immunological synapse 1
cell death
FC gamma R mediated phagocytosis
FC Epsilon RI signaling pathway
immunological synapse 2
T cell receptor signaling pathway
Cytokine and chemokine signaling
T cell homeostasis
NK cells pathway
induction of apoptosis 1
induction of apoptosis 2
negative regulation of apoptosis 3
MAPK signaling pathway 1
Macrophage-mediated immunity
wound healing
Regulation of NFkB transcription
NK cell mediated cytotoxicity
inflammation pathway
response to wounding
TGF Beta signaling pathway
MAPK signaling pathway 2
p38 MAPK signaling pathway 1
MAPK signaling pathway 2
FC Epsilon RI signaling pathway
P38MAPK pathway 1
P38MAPK pathway 2
p38 MAPK signaling pathway 2
Ras protein signal transduction
TGFβ receptor 1: an immune susceptibility gene in HPV-associated cancer

Chaya Levovitz, Dan Chen, Emma Ivansson, et al.

Cancer Res  Published OnlineFirst October 1, 2014.

Updated version  Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-14-0602-T

Supplementary Material  Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2014/10/02/0008-5472.CAN-14-0602-T.DC1

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.