Identification and characterization of HPV-host fusion transcripts in HNSCCs

Lada A. Koneva[1], Yanxiao Zhang[1,5], Shama Virani[1,2], Pelle B. Hall[1], Jonathan B. McHugh[4], Douglas B. Chepeha[3], Gregory T.Wolf[3], Thomas E. Carey[3], Laura S. Rozek[2,3,*], Maureen A. Sartor[1*].

[1]Department of Computational Medicine and Bioinformatics, [2]Department of Environmental Health Sciences, [3]Department of Otolaryngology/Head and Neck Surgery, [4]Department of Pathology University of Michigan; Ann Arbor, MI, USA. [5]Current address is Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, CA 92093.

Background Head and neck cancers together represent the sixth most common cancer worldwide. In 2015 the incidence of this type of cancer was estimated at > 742,000 new cases (> 400,000 deaths) [1]. The incidence of human papillomavirus (HPV)-related cancer in the upper aerodigestive tract has steadily increased over the past two decades, and now represents a majority of oropharyngeal cancer cases. Integration of the HPV genome into the host genome is a common event during carcinogenesis that has clinically-relevant effects if the viral early genes are transcribed. Understanding the impact of HPV integration on clinical outcomes of head and neck squamous cell carcinomas (HNSCCs) is critical for implementing deescalated treatment approaches for HPV-positive HNSCC patients.

Methods and Findings We identified and characterized integration events, represented by HPV-host fusion transcripts, using RNA-seq data from 84 HPV(+) HNSCC tumors: 18 patients from the University of Michigan Hospital (UM) and 66 from The Cancer Genome Atlas (TCGA). HPV integrations into the cancer genome were detected in 61% (51/84) of samples. Among the 18 HPV(+) UM tumors, viral-host fusion transcripts were found in nine (50%) of the samples. In the TCGA cohort we found 42 of the 66 tumors (63.6%) were integration-positive. Among the integration-positive tumors there were 41 HPV16 tumors, one HPV18, six HPV33, and three HPV35.

We found 320 virus-host fusion breakpoints, which were broadly distributed across the human and viral genomes and occurred within or near 89 human genes. Within the viral genome, breakpoints in oncogenes E6 and E7 were more common – 202 (63.13%) compared to breakpoints into other viral genes: E1 and E2 – 99 (30.94%), E4 and E5 – 44 (13.75%), or L1

and L2 – 15 (4.69%). This may be explained by preservation and expression of oncogenes E6 and E7 in all analyzed samples, while expression of the longest genes E1 and E2 were lost in more than half of the integration-positive tumors.

We investigated potential associations of integration sites with fragile and repetitive regions of the human genome, accounting for the regions of the genome covered by the RNA-seq data from all analyzed HPV(+) samples. We found significantly more insertional breakpoints in LINE, SINE, DNA elements, LTR and "All repeats" than expected by chance. There was no significant enrichment of HPV-host fusion breakpoints within common fragile sites (CFSs) although the p-value suggests a trend ($p = 0.058$). We did not find an enrichment of host-fusion breakpoints in non-fragile regions (NFR) ($p = 0.084$) or rare fragile sites (RFS), where fewer than expected by chance were identified.

Recurrent HPV integration events may signify the natural selection for tumor cells with breakpoints in specific genomic regions, and can suggest novel cancer driver genes. Of the 89 human genes with or near at least one identified integration event, five were associated with more than one tumor sample (i.e. recurrent integration). These genes were *CD274, FLJ37453, KLF12, RAD51B,* and *TTC6.* Genes with integration sites into exonic regions show elevated expression (OR = 11.6, Fisher's exact test p-value = 6.96×10^{-07}).

To further understand the biological context of genes associated with one or more insertional HPV breakpoints, we generated a protein interaction network directly connecting 65 of the 89 total genes (MetaCore software by Thomson Reuters). Within this resulting subnetwork, there were several hubs (genes with more than five interactions). These genes, in order from most to fewest interactions, were: *ETS2 (ETS), TP63, FOXA1 (HNF3), RUNX1 (AML1), KLF5,* and *CTGF (IGFBP7/8).* The network was highly statistically enriched for genes known to be important specifically in lung neoplasms ($p=1.69\times 10^{-26}$; rank=1), head and neck neoplasms ($p= 2.66 \times 10^{-11}$; rank=7), and urogenital neoplasms ($p=1.52\times 10^{-10}$; rank=9), suggesting selection for cells with integration events in key carcinogenic genes.

Using HNSCC overall survival data from TCGA, we found that patients with integration-negative tumors had better survival compared to those with integration-positive tumors (log-rank p-value = 0.04), which had a survival rate similar to patients with HPV(-) tumors (for three group comparison log-rank p-value = 0.0158). To investigate whether the difference in survival between integration-positive and integration-negative patients could be explained by differences

in biological processes, we performed enrichment analysis on the differentially expressed genes (DEGs) between the two groups. Differential expression analysis on all 84 HPV(+) samples using integration status as the group variable revealed 832 significantly DEGs (346 up in integration-positive and 486 up in integration-negative; FDR < 0.05 and |log2(FC)| > 1). Genes with elevated expression in integration-negative samples were most strongly enriched for immune related terms ("T cell activation", lymphocyte differentiation", "B cell activation" etc.), up-regulated genes in integration-positive tumors were enriched for keratinization and terms related to RNA metabolism and translation.

We hypothesized that enrichment of integration-negative samples for immune related genes could be explained by increased abundance of inflammatory cell types within these tumors. To test this hypothesis we used a cell-type-specific deconvolution technique to determine how the expression signatures of epithelial-relevant cell types differentiate the two groups. We used cell type specific signatures developed from a microarray database containing 723 samples associated with 25 epithelial-relevant cell types [2], and calculated a signature score across these 25 cell types for each of the 84 HPV(+) tumors. We found that integration-negative tumors had stronger immune signatures, characterized by heightened signatures for CD4+ T-cells, Regulatory T-cells, CD3+ T-cells, CD8+ T-cells, NK cells, NK T-cells, B cells, and CD34+ cells (Wilcoxon test; all p-values ≤ 0.01). They did not have significantly higher signatures for macrophages, gamma-delta T-cells, or neutrophils. The strongest cell type for integration-positive samples was keratinocytes ($p = 0.016$).

Conclusions In our study, we saw striking overrepresentation of integration events in or near genes known to be important to head and neck cancers, lung cancers, and urogenital cancers. Some of these genes were also recurrent and/or were hubs in our protein interaction network. Several lines of evidence point to the importance of *CD274 (PD-L1*) which is a members of the promising immune checkpoint pathways currently investigated in HNSCCs. Patients with no detected integration had better survival than those with a detected integration and HPV-negative patients. Our results suggest strong natural selection for tumor cells with expressed integration events in key carcinogenic genes.

Funding This work was funded by National Institutes of Health grants R01 CA158286, the University of Michigan Specialized Programs of Research Excellence (SPORE) grant (P50

CA097248), and the National Human Genome Research Institute (NHGRI) training grant (T32 HG00040).

1. Ferlay J, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. 2013; Available from: http://globocan.iarc.fr/Pages/burden_sel.aspx.
2. Swindell, W.R., et al., Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients. BMC Genomics, 2013. 14: p. 527.