SUPPLEMENTAL MATERIALS

Human papillomavirus and the landscape of secondary genetic alterations in oral cancers

Maura L. Gillison[1,*,+], Keiko Akagi[1,*], Weihong Xiao[1], Bo Jiang[1], Robert K. L. Pickard[2], Jingfeng Li[2], Benjamin J. Swanson[3], Amit D. Agrawal[4], Mark Zucker[5], Birgit Stache-Crain[6], Anne-Katrin Emde[7], Heather M. Geiger[7], Nicolas Robine[7], Kevin R. Coombes[5] and David E. Symer[8,*,+]

[1]Department of Thoracic/Head and Neck Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX
[2]Division of Medical Oncology, Department of Internal Medicine, Ohio State University, Columbus, OH
[3]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE
[4]Department of Otolaryngology – Head and Neck Surgery, Ohio State University Comprehensive Cancer Center, Columbus, OH
[5]Department of Biomedical Informatics, Ohio State University Comprehensive Cancer Center, Columbus, OH
[6]Complete Genomics, Mountain View, CA
[7]New York Genome Center, New York, NY
[8]Department of Lymphoma and Myeloma, University of Texas MD Anderson Cancer Center, Houston, TX

* these authors contributed equally to this work
+ Correspondence: Maura L. Gillison, M.D., Ph.D., mgillison@mdanderson.org;
David E. Symer, M.D., Ph.D., desymer@mdanderson.org

TABLE OF CONTENTS

SUPPLEMENTAL METHODS

Tumor samples in the Ohio cohort. In the Ohio cohort, fresh-frozen tumor was snap-frozen in liquid nitrogen within 30 minutes of resection, and DNA and RNA were purified after macrodissection using a cryostat to ensure ≥70% representation of tumor.

Genomic DNA sequencing. Genomic DNA was isolated from samples using standard methods involving phenol/chloroform extraction and isopropanol precipitation. Sample quality was assessed using a Nanodrop spectrophotometer and Picogreen double stranded DNA assay (Thermo Fisher Scientific).

From Ohio cohort samples, 59 T/N pairs were sequenced at ~90x mean depth of coverage by Complete Genomics WGS (CGI; **Supplemental Table S1**). The CGI aligner was used to map paired-end WGS reads (2 x 35 bp) against the human reference genome assembly GRCh37 (hg19) (Carnevali et al. 2012). The Ohio cohort also included 52 HPV-positive and 1 HPV-negative OSCC T/N pairs from which Illumina WGS data were generated at New York Genome Center (NYGC), including 40x mean coverage for normal samples and ~90x coverage for tumor. Illumina WGS data for 17 HPV-positive and 24 HPV-negative cancers were downloaded from TCGA. For Illumina data, sequence reads were aligned against human reference genome hg19 using BWA.aln version 0.5.9 (Li and Durbin 2010). We identified duplicate reads, realigned reads surrounding indels, and recalculated alignment quality scores using GATK v.3 (McKenna et al. 2010).

RNA-seq libraries and analysis. Total RNA was isolated from Ohio cohort OSCC samples using TRIzol (Invitrogen) extraction followed by isopropanol precipitation. Libraries were prepared using an Illumina TruSeq stranded RNA kit protocol resulting in ~350 nt cDNA inserts. Sequence data were generated using an Illumina HiSeq 2500 in high output mode. Paired end, 2 x 125 nt reads were sequenced at >40 million reads per sample. RNA-seq data for TCGA samples were downloaded from the TCGA portal (previously at https://tcga-data.nci.nih.gov/, migrated to https://portal.gdc.cancer.gov). RNA-seq reads from 18 HPV-positive and 2 HPV-negative OSCC Ohio cohort samples were aligned to GRCh37 human reference genome using STAR aligner 2.3.1z; 14 HPV-positive OSCC were aligned to GRCh37 using STAR 2.4.0c; and 52 HPV-positive and 24 HPV-negative OSCC were aligned to GRCh37.p13 reference assembly which includes non-canonical chromosomes (Dobin et al. 2013) and STAR 2.4.2a. RNA-seq reads from all TCGA samples were aligned using GRCh37.p13 and STAR 2.4.2a.

For batch correction of RNA-seq data, transcript structures first were downloaded from Gencode v.18 (http://www.gencodegenes.org/) as gene models to determine expression levels of each transcript. Aligned reads were counted using HTSeq (Anders et al. 2015), and raw counts were transformed into transcripts per million reads (TPM). Transcript expression values were adjusted for GC content using Bioconductor EDASeq (http://www.bioconductor.org/). Genes with low expression values (TPM < 1 in >80% of OSCC) were excluded from further analysis. A pseudo-count of one was added to TPM values to avoid undefined data upon log transformation. Resulting $\log_2$ TPM values were normalized and batch-corrected using the Bioconductor *sva* function ComBat (Leek et al. 2012). Variance in expression levels was calculated for each gene. Unsupervised hierarachical clustering of the 500 most highly variable genes in various defined sets of OSCC was performed using the Ward D2 method.

To assess the potential impact of distinct STAR aligner software versions and human genome reference assembly releases used in analysis of RNA-seq reads, principal components analysis was performed to assess RNA-seq sample data for 151 RNA-seq OSCC before and after reanalysis with the harmonized pipeline STAR 2.4.2a and reference genome assembly GRCh37.p13. The results revealed minimal impacts of the aligner and reference genome assembly version differences (**Supplemental Fig. S5K**).

HPV transcript analysis. To construct a custom, hybrid genome template for sequence alignments (Akagi et al. 2014), we downloaded the reference human (hg19) genome from UCSC genome browser (https://www.genome.org/) and concatenated it together with HPV types 16 (NC_001526.2), 18 (NC_001357.1), 31 (HQ537687.1), 33 (HQ537707.1), 35 (M74117.1) and 69 (AB027020.1) genomes from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/), and HPV types 52 (HPV52REF.1), 58 (HPV58REF.1), 56 (HPV56REF.1), 51 (HPV51REF.1), 59 (HPV59REF.1), 39 (HPV39REF.1) and 45 (HPV45REF.1) from PaVE database (https://pave.niaid.nih.gov/).

Somatic variant confirmation. Two approaches were taken to confirm somatic mutation calls. In the first, RNA-seq reads were aligned against the human hg19 reference genome using GSNAP (Wu and Nacu 2010). Quality and depth of mapped reads were evaluated, and mutation positions that were covered in RNA-seq data at > 20x coverage with uniquely aligned reads (mapping quality score [MAPQ] >=30) were identified. Mapped reads supporting alternative alleles comprising somatic variants were counted using samtools' mpileup (Li et al. 2009). We required the alternative alleles to be supported by >=5% of aligned RNA-seq reads. Mapped reads supporting alternative variant allele were counted using samtools' mpileup (Li et al. 2009) for each variant position.

In the second approach, a custom panel of Agilent SureSelect capture baits was designed for 61 recurrently mutated genes in HPV-positive OSCC and 58 genes in HPV-negative OSCC. These included the 24 and 25 significantly mutated genes identified by MutSig (Li et al. 2009) in HPV-positive and HPV-negative OSCC, respectively (**Fig. 1, Supplemental Fig. S1A**). Following hybrid capture of genomic DNA from 16 HPV-positive and 8 HPV-negative cancers, targeted resequencing was performed at average depth of sequencing coverage > 200x. Paired-end reads were aligned against human reference genome hg19 using BWA-MEM (Li 2013). Aligned reads were realigned and re-calibrated, and duplicate pairs were removed using GATK. Mapped reads supporting alternative variant allele were counted using samtools' mpileup. Variant calls supported by >40x depth of sequencing coverage, with >5% of mapped reads supporting the alternative allele, and showing no significant strand bias or map quality bias (each p<0.01), were considered to be confirmed (**Supplemental Table S1R**).

Comparison of human reference genome assemblies used in variant detection. To determine whether or not use of the more recent GRCh38 human genome reference assembly would substantially change the variant calls made with GRCh37 (hg19), as was used throughout this study, we re-analyzed WES data from 311 HPV-positive and HPV-negative T/N pairs. Out of 331 total WES samples analyzed in this study (**Supplemental Table S1F**), 311 samples were available for re-download from TCGA. We used BWA-0.7.15 to align sequence reads against hg19 and hg38, with the alternate locus-aware alignment protocol used for hg38. Mutect2 was used to call SNVs, based on the Mutect2 directory for the number of supporting reads for alternative alleles for each tumor/normal sample pair. Variants with population allele frequency > 0.01 and non-synonymous variants were identified using VEP v. 92. Results were displayed in a scatterplot comparing the $\log_{10}$-transformed variant frequencies per sample based on hg38 vs. hg19, and a correlation coefficient was calculated based on Pearson's correlation (**Supplemental Fig. S1K**).

To compare variant calls based on hg19 vs. hg38 for individual genes, WES data from 217 HPV-negative T/N matched pairs from TCGA were re-analyzed. These samples were comprised of an arbitrary subset of 285 HPV-negative T/N WES samples studied here (**Supplemental Table**

S1F). WES reads were aligned against hg19 and hg38 using BWA-0.7.15, with an alternate locus-aware alignment protocol used for hg38. As before, variants were called using Mutect2. Counts of somatic variants were determined for the 25 MutSig genes that were significantly mutated in HPV-negative OSCC, in comparing human reference genome assemblies hg19 vs. hg38. Variants were filtered for population allele frequency < 0.01 and for non-synonymous variants using VEP v. 92 (**Supplemental Table S3F**).

Annotation of somatic SNVs. To identify annotation identifiers for somatic SNVs, their chromosomal coordinates and nucleotide substitutions as mapped to the hg19 reference genome assembly were submitted as inputs to VEP v. 94 at Ensembl to query the COSMIC (https://cancer.sanger.ac.uk/cosmic), dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/) and Ensembl variation databases. Variant ID outputs were refined further by manual inspection (**Supplemental Table S4A**).

Gene ontology. Gene ontologies and statistical significance of enrichment were assessed using the Panther database and suite of analytical tools (www.pantherdb.org/) (Mi et al. 2013).

CNV detection. Mapped WGS sequence reads from each T/N pair for 103 HPV-positive OSCC and 50 HPV-negative OSCC were counted in 2 kb bins genome-wide. For CGI WGS data, mean depths of coverage were extracted for reads mapped to each 2 kb genomic bin from coverageRefScore data generated by the CGI data analysis pipeline. For Illumina WGS data, paired-end reads were aligned using BWA.aln (Li and Durbin 2010) and processed using GATK v3. To adjust for differences in depths of coverage within or between paired T/N samples, reads were down-sampled to 100 million per sample, using DownsampleSam in Picard tools (Broad Institute). To mitigate differences between platforms and alignment protocols, only uniquely aligned Illumina reads with reliable alignment scores (MAPQ >= 30) were counted. For both CGI and Illumina data, the Bioconductor package CNAnorm was used to correct genomic G/C nucleotide content, normalize read counts and detect somatic copy number alterations (Gusnanto et al. 2012). The relative ploidy of each 2 kb bin was calculated as the ratio of sequencing depth of coverage in each tumor vs. its matched normal sample. These ratios were smoothed using the DNAcopy algorithm within CNAnorm (Venkatraman and Olshen 2007). DNA copy number status was segmented by merging adjacent 2 kb bins based on similar average ratios of copy numbers. In subsequent analysis, genomic segments (based on 2 kb bin resolution) having a copy number (i.e. estimated ploidy) < 1.5 were defined as regions of copy number loss, while segments with copy number > 2.5 were defined as regions of copy number gain. For analysis of gene-level copy number in samples studied by TCGA WES, segmented copy number calls (based on microarray data) were downloaded from the TCGA portal.

Frequency distribution of copy number gains or losses. To calculate the expected number of copy number alterations for each tumor, amplified segments (resulting from genomic segmentation as described above) were permuted 100 times. Similarly, lost segments also were permuted for each tumor. We used the binomial distribution to calculate the significance of the observed vs. expected number of samples harboring amplified segments anywhere within 500 kb bins, and those bins with $p<0.01$ were considered to be significantly amplified. Comparable analysis identified 500 kb bins with significant copy number losses. Cumulative lengths of significantly gained or lost 500 kb bins were calculated for each chromosome arm as displayed (**Fig. 5**). The number and fraction of OSCC samples with copy number estimates were counted and plotted for each bin. In a second approach, we used GISTIC 2.0.22 (Mermel et al. 2011) to identify significant alterations in somatic copy numbers of broad genomic regions, and obtained similar results (data not shown).

REFERENCES

Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, Rocco JW, Teknos TN, Kumar B, Wangsa D et al. 2014. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res* **24**: 185-199.

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.

Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB, Pant KP, Ebert JC, Brownley A, Morenzoni M, Karpinchyk V et al. 2012. Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* **19**: 279-292.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28**: 40-47.

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**: 882-883.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**: R41.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8**: 1551-1566.

Venkatraman ES, Olshen AB. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**: 657-663.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873-881.