**Genomic Landscape of Head and Neck Cancer in Asia: A Comprehensive Meta-Analysis of 1016 samples**

Srikanth S. Manda1,2, Nafisa Arfa3, Neha Sharma1,2, Aparna R. Parikh4, Thomas J. Roberts4, Sewanti Limaye5, Venkataraman Ramachandran1,2,6, Kumar Prabhash7,8, Moni A. Kuriakose1,2, Prashant Kumar1,2,6,#

**Supplementary methods**

**Recurrence gene mutation criteria**

A gene mutated in two or more samples was considered as recurrently mutated in that cohort. Furthermore, gene mutations occurring in at least two samples in at least two countries were referred to as recurrently mutated genes in the continent cohort. All relevant bioinformatic analysis was done on the recurrently mutated genes in the respective cohorts.

**Cancer Census Genes**

The gene mutations were investigated for their involvement in carcinogenesis using the Catalogue of Somatic Mutations in Cancer COSMIC (<https://cancer.sanger.ac.uk/cosmic/browse/tissue>)  Cancer Gene Census database which lays down an ensemble of mutations in genes driving tumorigenesis across all cancers. Querying in the Cancer Browser of COSMIC, in the head neck section under the upper aerodigestive tract, gives an exhaustive list of 719 out of the 737 cancer driver genes found in HNSCC samples. The CGC genes cancer driver genes were identified among the whole set of genes from the extensive list of 737 cancer driver genes reported in the database.

**TCGA data**

The data pertaining to the gene mutation frequency in The Cancer Genome Atlas (TCGA) Program were procured from the web-based browser of cbioportal (<https://www.cbioportal.org/>). Querying the TCGA studies (TCGA Firehose Legacy, TCGA Nature 2015, and TCGA PanCancer Atlas) among the 15 head and neck cancer-related studies provides genes with frequency across 1332 samples which was subsequently used for comparison with the frequencies in the Asian cohort.

**Visualisation of the mutation profile**

The visualization of the mutation profile using heatmaps was done using R-packages v.4.3.1. The oncoplots of genes with frequency were generated using the Maftools R package [9]. The mutually exclusive or co-occurring gene mutation was explored using the Somatic Interactions feature in Maftools that uses a Pair-wise Fisher's Exact test. P-value< 0.05 was considered significant.

**Driver Gene Analysis**

The gene mutations across the study cohort comprising 425 samples from the ten Whole Exome studies were examined for potential driver genes using OncodriveFML v.2.2.0 [10](<https://bbglab.irbbarcelona.org/oncodrivefml/home>), a novel tool for detecting putative cancer driver genes. The input file for this analysis comprised the whole set of 14,655 genes with variant information obtained exclusively from the whole exome datasets. A p-value cutoff of <= 0.05 was applied to obtain statistically significant potential cancer driver genes. The statistically significant genes are then inspected for their frequencies in the sample cohort and sorted based on the same.

**Pathway enrichment analysis**

Pathway enrichment analysis was performed on recurrently mutated genes using Enrichr (<https://maayanlab.cloud/Enrichr/>), a web-based gene set enrichment analysis tool. The 500 most frequently mutated genes were analysed for oncogenic pathways in the Molecular signatures database (MSigDB) database [11]. The analysis criteria for identifying statistically significant altered signaling pathways included a p-value cutoff <= 0.05 and a minimum of 3 genes in each pathway. The enriched pathways were further checked for the fraction of samples and pathways affected using the OncogenicPathways feature in the Maftools R package. The interaction network among the genes involved in a pathway was elucidated using the STRING functional protein association network.

**Survival plots**

The Kaplan-Meir survival plots were obtained from cBioportal, a web-based resource for the visualization of multidimensional cancer genomic datasets. The gene of interest was queried for survival data in HNSCC from the portal. KM plots were downloaded and patients segregated with mutant and wild type version of the gene. The log-rank test of difference was used to determine the p-value. A p-value <= 0.05 was considered statistically significant for further analysis.

**Comparison plots**

The available clinical factors from various studies with complete data were used for the comparative analysis. The cohort comparison primarily comprising smokers and non-smokers was done using functions provided in maftools R package. The forest plot delineating the differentially mutated genes in both cohorts was procured using forestPlot function with minimum mutated sample criteria set as 7. The co-bar plot elucidating the mutation frequency difference across the cohorts was furnished using the coBarplot function that depicts genes that qualified the following criteria: CGC gene present among the top 20 most frequently mutated genes in both the cohorts; potential cancer driver genes, as identified in the driver gene analysis, among the top 20 most frequently mutated genes in both the cohorts; and CGC genes with a difference greater than 5% across the cohort. The statistically significant (p-value <0.05) putative oncogenic signaling pathways were identified using pathway enrichment analysis on the top 500 most frequently mutated genes in both cohorts using the same criteria like the overall pathway analysis. The differences in enriched pathways were delineated through a co-bar plot with relevant pathway characteristics evaluated using Maftools R Package.