JCI insight

RESEARCH ARTICLE

Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis

Wenjin Liu,[1,2,3] Jeff M. Snell,[3,4,5] William R. Jeck,[1,3] Katherine A. Hoadley,[1,3] Matthew D. Wilkerson,[1,3] Joel S. Parker,[1,3] Nirali Patel,[3,6] Yohannie B. Mlombe,[7] Gift Mulima,[8] N. George Liomba,[9] Lindsey L. Wolf,[9,10] Carol G. Shores,[3,11] Satish Gopal,[2,3,7,9,12] and Norman E. Sharpless[1,2,3]

[1]Department of Genetics, [2]Department of Medicine, [3]The Lineberger Comprehensive Cancer Center, [4]Program in Bioinformatics and Computational Biology, [5]Program in Molecular and Cellular Biophysics, [6]Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, North Carolina, USA. [7]Department of Medicine, University of Malawi College of Medicine, Blantyre, Malawi. [8]Department of Surgery, Kamuzu Central Hospital, Lilongwe, Malawi. [9]UNC Project-Malawi, Lilongwe, Malawi. [10]Department of Surgery, Brigham and Women's Hospital, Boston, Massachusetts, USA. [11]Department of Otolaryngology/Head and Neck Surgery, [12]Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, USA.

Esophageal squamous cell carcinoma (ESCC) is endemic in regions of sub-Saharan Africa (SSA), where it is the third most common cancer. Here, we describe whole-exome tumor/normal sequencing and RNA transcriptomic analysis of 59 patients with ESCC in Malawi. We observed similar genetic aberrations as reported in Asian and North American cohorts, including mutations of *TP53*, *CDKN2A*, *NFE2L2*, *CHEK2*, *NOTCH1*, *FAT1*, and *FBXW7*. Analyses for nonhuman sequences did not reveal evidence for infection with HPV or other occult pathogens. Mutational signature analysis revealed common signatures associated with aging, cytidine deaminase activity (APOBEC), and a third signature of unknown origin, but signatures of inhaled tobacco use, aflatoxin and mismatch repair were notably absent. Based on RNA expression analysis, ESCC could be divided into 3 distinct subtypes, which were distinguished by their expression of cell cycle and neural transcripts. This study demonstrates discrete subtypes of ESCC in SSA, and suggests that the endemic nature of this disease reflects exposure to a carcinogen other than tobacco and oncogenic viruses.

Introduction

Esophageal squamous cell carcinoma (ESCC) is the sixth most common cause of cancer death worldwide (1, 2). Endemic ESCC occurs in specific regions of the world including 2 high-risk belts in central Asia and sub-Saharan Africa (SSA), where the incidence is more than 20-fold higher than in Western countries. In Malawi, a country of nearly 17 million people in southern Africa, ESCC is the third most common cancer, accounting for 12% of all malignancies (3). At Kamuzu Central Hospital in Lilongwe, Malawi, ESCC was diagnosed in 27% of esophagogastroduodenoscopies performed (4). ESCC patients in SSA are typically diagnosed at advanced stages with substantial dysphagia and malnutrition, and suffer poor outcomes (5).

The underlying reasons for high ESCC incidence in SSA remain unclear. While tobacco exposure and alcohol are considered the main risk factors in Western countries, they are not major factors in high-incidence areas, and in most studies greater than 60% of cases are nonsmokers (6). The high incidence of several cancers in SSA including Kaposi's sarcoma, hepatocellular carcinoma, Burkitt's lymphoma, and cervical cancer are known to be secondary to the oncogenic viruses human herpesvirus 8 (HHV8), hepatitis B virus (HBV), Epstein-Barr virus (EBV), and human papillomavirus (HPV), respectively. However, evidence for a novel or established oncovirus etiology in ESCC has yet to be identified. Prior work in Brazil, China, and Iran has suggested that ingestion of polycyclic aromatic hydrocarbons through traditional teas and other sources, thermal injury from scalding hot beverages, and dietary selenium deficiency may be potential etiologic contributors (7–11). However, definitive studies examining these and other risk factors in SSA have not been performed. Recent studies suggest that ESCC is more common in rural areas of Malawi where people are more likely to use wood for cooking, grow their own maize and store it in sacks at home, and use untreated water sources (4, 6).

Authorship note: W. Liu and J.M. Snell contributed equally to this work.

Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: June 21, 2016
Accepted: September 7, 2016
Published: October 6, 2016

Reference information:
JCI Insight. 2016;1(16):e88755.
doi:10.1172/jci.insight.88755.

Table 1. Patient characteristics and demographics

Feature	Category	Malawi cohort, *n* = 59
Age (years)	Mean	56
	Range	24–93
Sex [*n* (%)]	Male	27 (45.8)
	Female	31 (52.5)
	Unknown	1 (1.7)
Smoking [*n* (%)]	Yes/Former/ Chingambwe	24 (40.7)
	No	33 (55.9)
	Unknown	2 (3.4)
Alcohol [*n* (%)]	Yes	14 (23.7)
	No	18 (30.5)
	Unknown	27 (45.8)
HIV [*n* (%)]	Negative	44 (74.6)
	Positive	10 (16.9)
	Unknown	5 (8.5)
Tumor differentiation [*n* (%)]	Well	7 (11.9)
	Moderately	17 (28.8)
	Poorly	6 (10.2)
	Unknown	29 (49.1)

While the Cancer Genome Atlas (TCGA) and related efforts have comprehensively analyzed the most common malignancies in Western countries, ESCC has been relatively understudied. While adenocarcinoma of the proximal stomach and esophagus is becoming more common in the United States and Europe, the incidence of ESCC in Western nations has been in decline for decades (12, 13). Moreover, in the United States, ESCC is generally treated with neoadjuvant chemoradiotherapy, making the acquisition of pristine tumor tissue for comprehensive molecular analysis challenging. In 2 other endemic regions, China and Japan, molecular studies have described the mutational landscape for ESCC in those countries (14–17). These analyses demonstrate frequent mutations in genes associated with other squamous cancers, including *TP53*, *RB1*, *CDKN2A*, *PIK3CA*, *NOTCH1*, and *NFE2L2*, as well as other potentially novel cancer-associated genes and histone modifier genes.

Unlike higher-resource settings, molecular biology approaches have rarely been applied to cancer in SSA. Throughout the region, there is limited pathology infrastructure, tissue procurement capability, and laboratory capacity to conduct genomic studies, as well as legitimate cultural and regulatory sensitivities regarding germline and somatic genomic research and transfer of biologic specimens outside SSA. For ESCC, obtaining tumor specimens is especially challenging in settings with scarce endoscopy and thoracic surgery facilities. In Malawi, ESCC patients often do not undergo endoscopy or biopsy even when services are possible, due to clinicians' sense of futility when palliative stenting cannot be simultaneously performed, in a setting where stents are costly and typically unavailable.

Given the high incidence of ESCC in SSA and the marked differences in the epidemiology of ESCC in the West and Asia versus SSA, we undertook a dedicated analysis of African tumors. To that end, 59 untreated ESCCs and matched normal pairs were analyzed from consenting Malawian patients with ESCC, participating in a case-control study conducted at 2 national teaching hospitals in Lilongwe and Blantyre from 2011 to early 2013. Whole-exome DNA sequencing was performed on tumor and paired normal DNA, as well as whole-transcriptome RNA sequencing. Somatic mutations and copy number events were identified using approaches developed for TCGA (18). This analysis provides a comprehensive molecular evaluation of ESCC, an understudied and highly lethal cancer that poses a significant global public health burden.

Results

Description of the patient cohort. 59 patients with ESCC participating in a case-control study at Kamuzu Central Hospital, Lilongwe, Malawi and Queen Elizabeth Central Hospital, Blantyre, Malawi were enrolled in this study throughout 2011 and 2012. Potential participants were initially identified by clinical symptoms including dysphagia, odynophagia, hematemesis, and weight loss. ESCC diagnosis was confirmed by histopathological sections after endoscopic biopsy or rarely surgical resection. All pathology underwent central re-review at the University of North Carolina. Patients over 18 years of age and willing to participate in this study provided written informed consent in their native language and had matched normal tissue taken concurrently at time of biopsy or resection. Human studies approval was given by the Malawi National Health Sciences Research Committee and the University of North Carolina Internal Review Board.

The study population included 27 males (46%) and 31 females (53%) (Table 1). In accord with the age distribution of ESCC in SSA, this cohort was younger than cohorts in other continents, with the average age being 56 (range, 24–93) years and with 17 (29%) patients younger than 45 years old. Life expectancy in Malawi during the study period was approximately 55 years with 5% of the population aged 60 years or older, suggesting that our ESCC cohort was older than the Malawi general population (19). Roughly 18% of the Malawian population reports tobacco use including local cigarettes and Chingambwe (20), a chewed local tobacco, whereas 41% (24 of 59) of the ESCC cohort reported tobacco use, suggesting a modest enrichment for tobacco users in this population, but the majority (56%) of the cohort were never smokers. Of patients with known HIV status, 10 (17%) of the ESCC population were HIV^+, which is slightly higher than the 11% population prevalence of HIV during the study period (21). Cooking method and maize

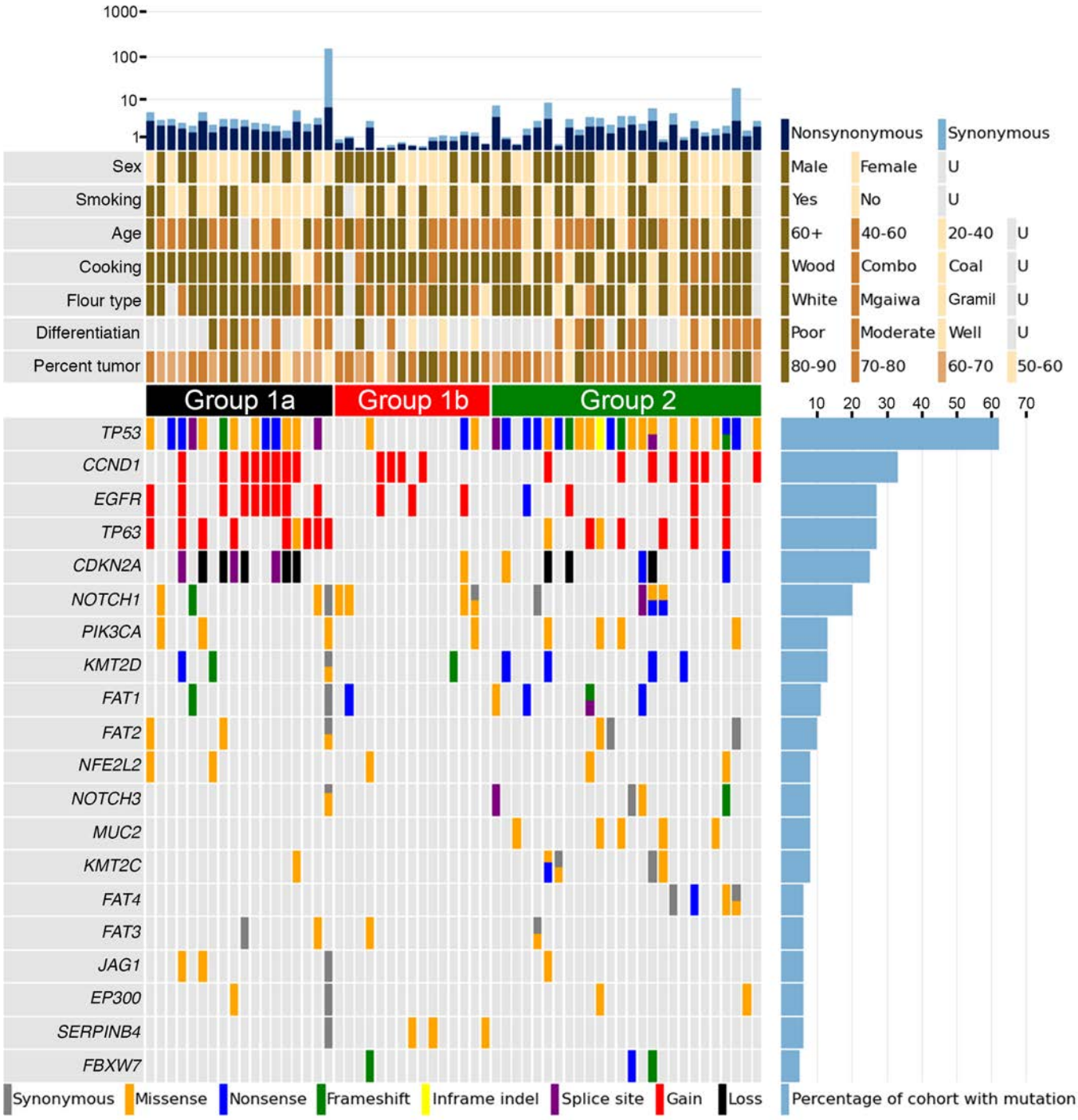

Figure 1. Summary of patient characteristics and somatic point mutations. Each column represents 1 of 59 patient samples and columns are ordered based on RNA subtype (see Figure 4). The frequency of synonymous and nonsynonymous mutations per Mb sequenced (log scale) are stacked and shown at the top of the figure. Clinical and environmental characteristics of each sample are displayed in a matrix in the center of the figure, where each row is a single characteristic. Somatic single-nucleotide polymorphisms and copy number variations of each sample are displayed in a matrix in the bottom of the figure, where each row is a single mutated gene ordered top-to-bottom by decreasing percentage of the cohort with mutations in a given gene. Each mutation type is color coded and samples with 2 or more mutation types per gene are indicated by 2 or more colors. Percentage of the cohort with mutations in a given gene are shown in the bottom right of the figure.

flour type, both found to correlate with ESCC in this population (6), are indicated in Figure 1. Maize is the dietary staple of Malawi and is processed in various ways. M'gaiwa is milled whole-grain maize, while gramil is dehulled, then milled. White maize is dehulled, soaked, washed, dried, and then milled, and likely contains the fewest contaminants of the 3 flours.

Table 2. *TP53*-specific analysis of hotspots at sites of known mutations among patients with *TP53* SNPs

Carcinogen	Hotspots	Malawi cohort, *n* = 47
Smoking [*n* (%)]	157, 158, 273	3 (6.3)
Aristolochic acids [*n* (%)]	131, 209, 280	0 (0.0)
Aflatoxin B1 [*n* (%)]	249	0 (0.0)
UV radiation [*n* (%)]	248, 278	2 (4.3)

Mutations and copy number events of the tumors. We sequenced the exomes of 59 untreated ESCCs and matched normal pairs at a mean sequencing coverage of 85×, with 97% of targeted bases above 30× coverage. We simultaneously performed RNAseq on these same 59 tumors at an average of 150 million reads per sample. Through a combined analysis of both RNA and DNA sequencing (22) we identified between less than 0.1 and 20 mutations per Mb of sequenced target, with the exception of 1 sample, which exhibited mutations of genes involved in mismatch repair, demonstrating 156 mutations per Mb (Figure 1). Mutations found at 5% or greater frequency are shown in Figure 1. In common with prior studies of head and neck squamous cancers in the US (23) and ESCC in Asian populations (14–17), we found a significant enrichment of inactivating or dominant-negative events of the known tumor suppressor genes *TP53*, *CDKN2A*, *NOTCH1/3*, *FAT1/2/3/4*, and *FBXW7*, as well as known activating events of *PIK3CA* and *NFE2L2*. There was an enrichment of mutations in genes encoding the chromatin-modifying enzymes *KMT2D* (*MLL2*), *KMT2C* (*MLL3*), and *EP300*. Mutations of unclear biochemical effect were enriched in 2 other genes indirectly associated with squamous cancers (24, 25): *JAG1*, the NOTCH ligand, and *SERPINB4*, a granzyme inhibitor.

We queried for areas of significant copy number change through genomic identification of significant targets in cancer (GISTIC) analysis (Figure 2, with samples ordered as per Figure 1). We identified statistically enriched copy gains in *TP63*, *MYC*, *CCND1*, *ERBB2*, *CCNE1*, and *MYCL1*. Regions of significant copy losses included 1q31, 3p, 5q, 9p21.3 (*CDKN2A/CDKN2B*), 10q25, and 21q22. Except for the deletions noted at 1q31, these regions of amplification or deletion have been reported in Asian ESCCs and other squamous cancers (14–17). The 1q31 deletion was focused around *CDC73*, an established suppressor of parathyroid neoplasia. High-prevalence point mutations and copy number events are summarized in Figure 1. The spectrum of point mutations and copy number alterations (CNAs) did not significantly correlate with measured epidemiologic exposures, patient clinical characteristics, or histologic features of the disease (Figure 1). We observed an association of CNAs with RNA subtype, with subtypes 1a and 2 having more CNAs than subtype 1b (Figure 2, and see below). These data in aggregate suggest that the genes targeted for point mutation and CNAs in African ESCCs are similar to those mutated in Asian ESCCs as well as squamous malignancies of the lung, head, and neck in Western populations.

Mutational spectral analysis and pathogen detection. As ESCC prevalence in some populations has been tied to carcinogen exposure, we sought to identify potential causes of ESCC by mutation signature analysis (26, 27). We performed an analysis on this cohort of 59 ESCCs as well as on published exome data of 139 ESCCs from a Chinese population (15). Stability analysis revealed 3 mutation signatures in SSA ESCCs (Figure 3A), as opposed to only 2 stable signatures in Chinese ESCCs (Figure 3B). Two signatures were consistent between SSA and Chinese cohorts, and matched previously reported patterns. The first shared signature, with abundant C to T transition mutations, is consistent with the age-associated signature seen in numerous types of cancer. The second, with both C to G transversions and C to T transitions occurring at TCN trinucleotides, closely matches the APOBEC-associated cytidine deaminase signature, common in many malignancies (28).

In our SSA cohort, however, a third signature was observed that lacked overlap with prior signature analyses, including UV light, mismatch repair, DNA polymerase ε, aristolochic acid, aflatoxin, or BRCA1/2 mutation signatures. This third unknown signature was similar to the previously reported signature 29, which was associated with gingival buccal oral squamous cell carcinoma from individuals known to chew tobacco (29). It is characterized by an excess of C to A transversions (predominantly at NCA trinucleotides) and C to T transitions (predominantly at NCG trinucleotides). Concordant with these signature analyses of whole-exome data, a dedicated analysis of *TP53* mutation hotspots did not identify known carcinogen-associated hotspot mutations (Table 2). Together, these results suggest that SSA ESCC is induced by mutations caused by age-associated spontaneous deamination of 5-methyl-cytosine, APOBEC-induced deamination of cytosines that follow thymidines, and an occult mutagenic process that nonrandomly targets either cytosines or guanines to induce transitions and transversions. Importantly, we did not note evidence in either whole-exome analyses or *TP53* mutations for excess mutagenic events related to several

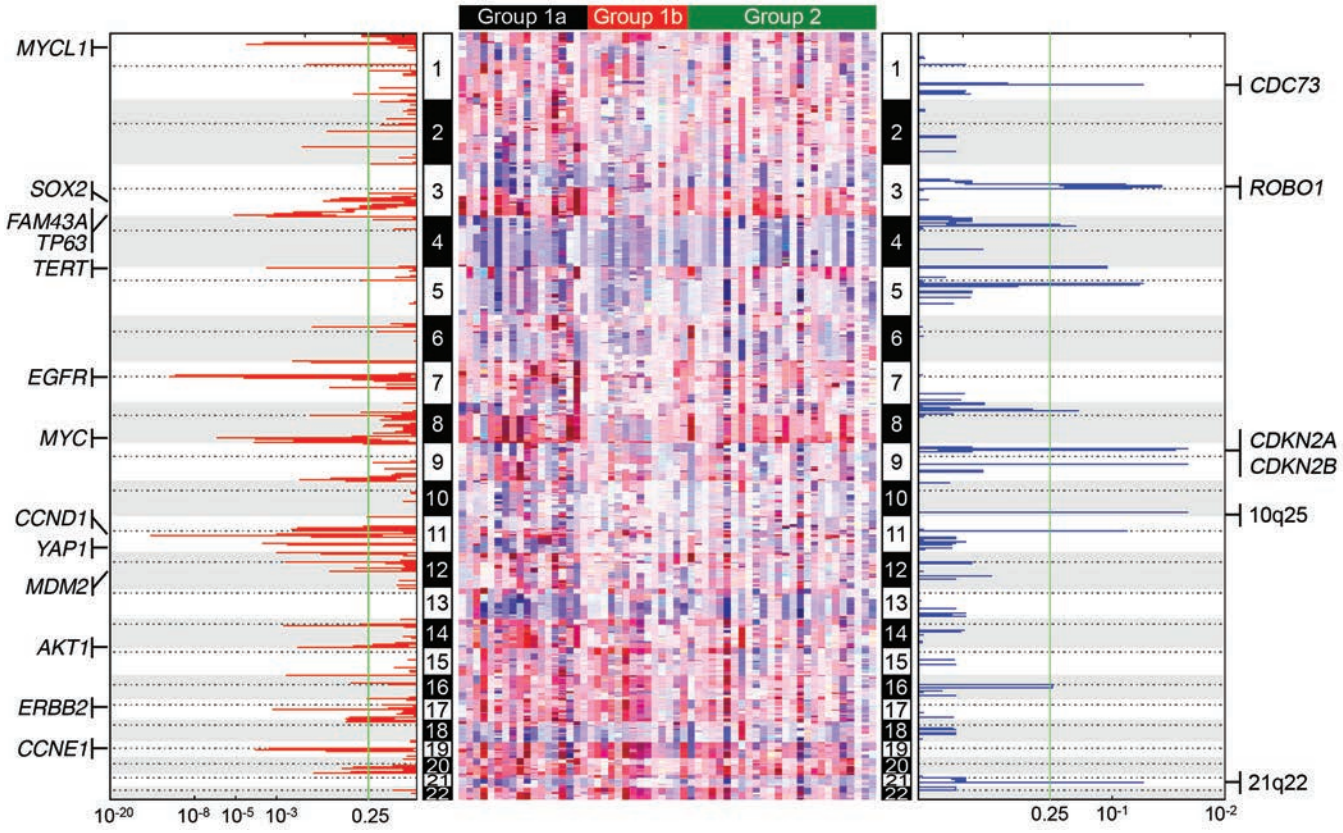

Figure 2. Summary of copy number alterations (CNAs). Genomic identification of significant targets in cancer (GISTIC) analysis of 59 esophageal squamous cell carcinoma (ESCC) samples for significantly amplified or deleted regions. The genomic profile of the copy number data is displayed in a heat map in the center of the figure, where samples are ordered based on RNA subtype (see Figure 4). Significant amplifications are indicated on the left in red and deletions on the right in blue. The score at the bottom is the false discovery rate *q* value as calculated by GISTIC.

well-known environmental carcinogens (e.g., cigarette smoking, aflatoxin, mismatch repair). Additionally, we did not identify mutational signatures unique to groups exposed to previously reported potential local carcinogens (e.g., white maize flour, wood-fueled cooking) (6).

We considered the possibility that an occult infectious pathogen contributes to the etiology of SSA ESCC. In particular, EBV has been associated with nonsquamous gastric cancer and Burkitt lymphoma (endemic in Malawi) and HPV has been associated with squamous malignancies of the cervix, anus, and oropharynx, with prior reports in conflict regarding the role of HPV in ESCC (30–33). We therefore looked for an occult pathogen in 2 ways: we searched RNAseq data for reads with homology to known viral transcripts, and also assembled unmapped DNAseq reads and searched these unmapped contigs for homology to databases of infectious agents. Using the former approach, we readily identified evidence of chronic herpesviruses (e.g., EBV) and HIV in a small minority of the samples. As these viruses are commonly found in circulating hematopoietic elements in humans, we believe these findings represent contamination of the tumor biopsies with virus-containing lymphocytes rather than evidence for a tumor-promoting role of these events. Consistent with this view, the read count for these viral sequences was low, and was seen in a proportion of samples that was lower than the expected prevalence for the Malawian population (e.g., 11% for HIV and greater than 90% for EBV) (21, 34–36). Importantly, neither the RNA nor DNA analysis suggested HPV infection, whereas these algorithms readily detect HPV sequences in HPV-infected cervical or oropharyngeal cancers. Further evidence for a lack of HPV infection was the high frequency of *TP53* and *CDKN2a* inactivation, which are uncommon in HPV-induced tumors that express viral oncoproteins targeting the p53 and retinoblastoma (RB) pathways. Although these analyses cannot exclude the possibility that a minority of SSA ESCCs are caused by an occult pathogen or by a pathogen that is not present in malignant cells (e.g., as with *Helicobacter pylori* and gastric cancer), these results suggest that SSA ESCC is not caused by an integrated (e.g., HPV) or episomal (e.g., EBV) tumor virus.

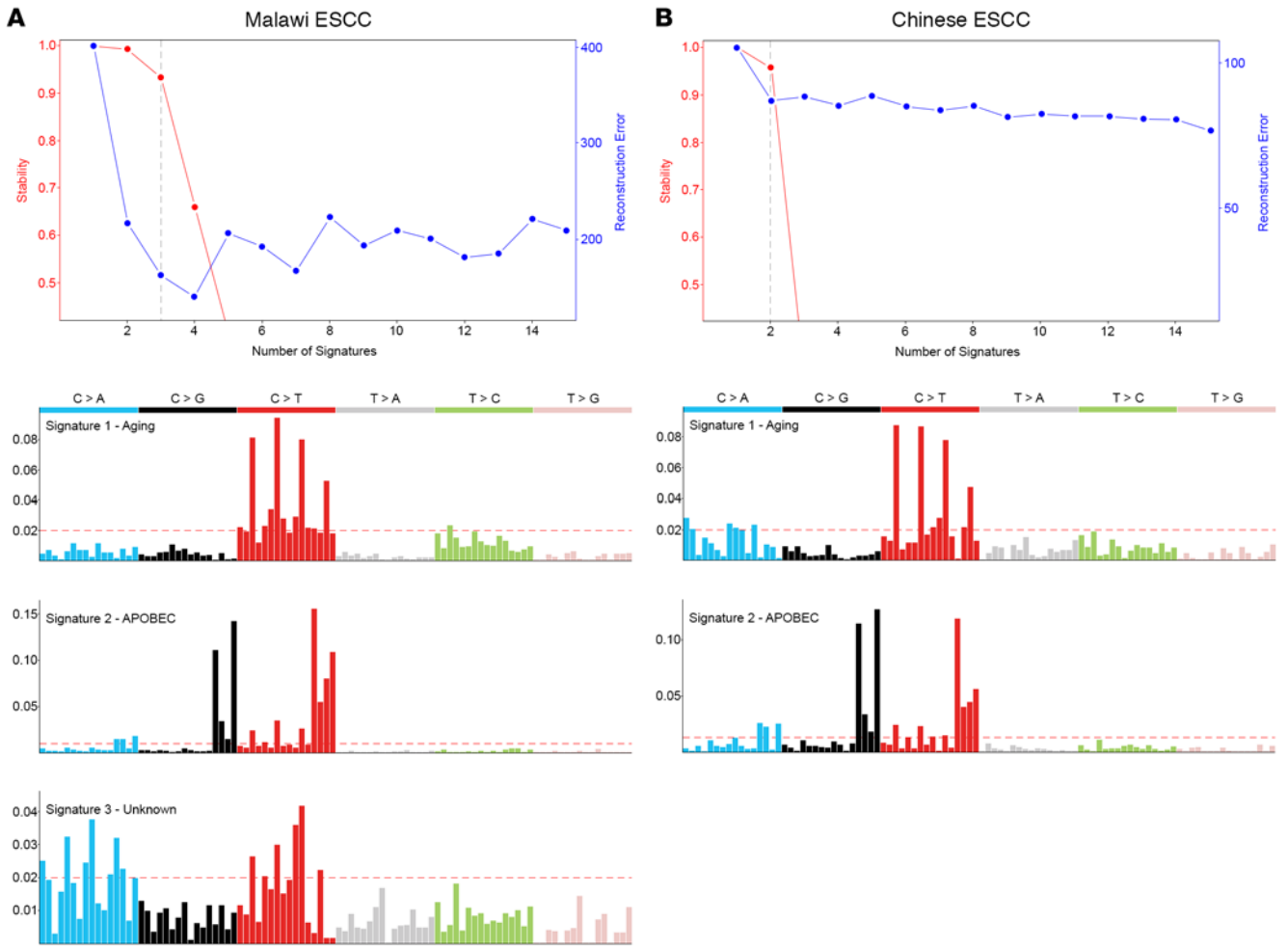

Figure 3. Mutational analysis of exome data and *p53*. Signature stability (top) and mutational signatures (bottom) are shown for Malawian (**A**) and Chinese (**B**) exome analyses. Stability is a measure of the reproducibility of a signature and reconstruction error is a measure of the reproducibility of the original mutation catalog (25). In Chinese patients, 2 signatures (aging and apolipoprotein B mRNA editing enzyme catalytic polypeptide-like [APOBEC]) are noted, whereas in Malawian esophageal squamous cell carcinoma (ESCC), a third signature is also identified, characterized by cytosine to adenine transversions and cytosine to thymine transitions.

RNA expression analysis. We next considered whether samples in our cohort could be divided into gene expression subtypes. We first validated our RNAseq results by performing hierarchical clustering of our data combined with TCGA RNAseq datasets from the PanCan analysis (18). All of the ESCC samples clustered together, and nearest to squamous carcinomas of the head, neck, and lung. Lung adenocarcinoma and other nonsquamous tumors were more distant from ESCC (Supplemental Figure 1; supplemental material available online with this article; doi:10.1172/jci.insight.88755DS1). SSA ESCC, like other squamous tumors, was characterized by increased expression of transcripts such as *TP63*, squamous cytokeratins (e.g., *KRT5/15*) and keratinocyte-specific transcripts (e.g., *BNC1*, *DSC3*, and *DSG3*) This, in addition to our careful histologic re-review of all samples, supports the correct diagnosis of these tumors as squamous malignancies.

We next performed hierarchical clustering using dynamic genes quantified in these ESCC primaries, showing 3 distinct subtypes of SSA ESCC. Gene set analysis of the subtypes revealed differences most prominently in their expression of markers for neural differentiation, morphogenesis, and DNA repair and metabolism (Figure 4 and Supplemental Figures 2–5). We performed this analysis independently on 2 split subgroups of the samples, and saw evidence for these subtypes in both analyses, suggesting that this finding is robust. Subtype 1 could further be divided into 2 subgroups (1a and 1b) of nearly equal size (18 and 15 samples, respectively), which differed further in expression of transcripts associated with DNA replication and repair, small GTPases, and homeobox genes (e.g., *HOXA11/C11* and *SIX1/4*). Silhouette width analysis and principal component analysis also supported the existence of 3 distinct subtypes of SSA ESCC

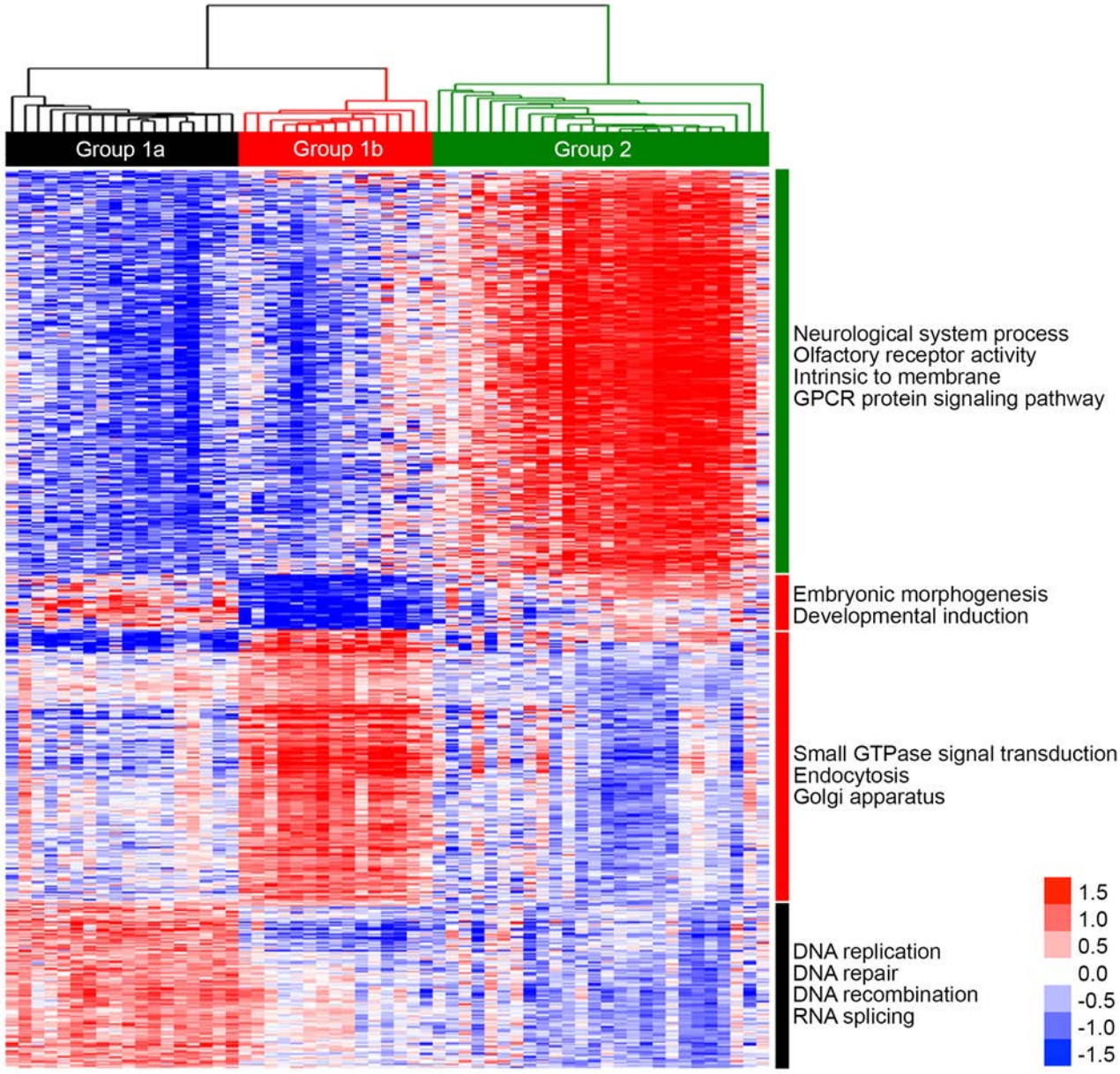

Figure 4. RNA expression analysis of Malawian esophageal squamous cell carcinoma (ESCC) tumors. A heat map of transcripts identified as being dynamically expressed (red, upregulated transcripts; blue, downregulated transcripts) among the 59 patient samples. Unsupervised hierarchical clustering suggests 2 subtypes (1 and 2) with significantly different expression profiles. Subtype 1 can be further divided into 2 subtypes (1a and 1b) that are distinguished by frequency of copy number alterations (CNAs), point mutations, and *TP53* mutations (Figures 1 and 2). The finding of 3 subtypes by RNA expression analysis is consistent with silhouette width and principal component analysis (Supplemental Figures 6 and 7).

based on RNA expression (Supplemental Figures 6 and 7). Patients of subgroup 1b were more likely to be smokers, more likely to be male, more likely to report use of wood-burning cooking methods, less likely to report use of more refined white corn flour, and were older (Figure 1).

We searched for defining features of these potentially novel subtypes. Subgroup 1b was more quiet genomically, with the fewest somatic mutations per Mb (0.75) compared with subgroup 1a (11.85) and subtype 2 (3.71) (Figure 1). Subgroup 1b also demonstrated fewer CNAs (Figure 2), most notably with fewer amplifications of *MYC*, *EGFR*, and *TP63* and fewer deletions of *CDKN2A/B* compared with other subgroups. Along with these differences in genomic complexity, subgroup 1b also was significantly less likely to exhibit *TP53* mutation ($P < 0.05$, ANOVA), in accord with p53's known role in regulating DNA

repair and ploidy. Mutations of a few other genes (e.g., *MUC2*, *SERPINB4*) appeared to be enriched in specific subtypes, but given our total sample size, our study was underpowered to characterize the molecular landscapes of each subtype. In aggregate, these observations support the existence of 3 distinct molecular subtypes of SSA ESCC that differ in clinical features, genomic complexity, p53 mutational status, and RNA expression.

Discussion

In this work, we performed integrated genomic analysis of SSA ESCC. This includes dedicated, unbiased analyses for somatic DNA mutations of the exome, CNAs, RNA expression profiling, potentially novel mutagenic signatures, and occult pathogens. Our work has shown that in molecular terms, the genomic events (driver activation and tumor suppressor gene loss) that cause ESCC in Malawi appear to be similar to those in the Western countries Japan and China. Our data also suggest, however, that the mechanism of mutagenesis, that is, the epidemiologic cause of this highly common cancer, is likely different between ESCC in SSA versus ESCC in other parts of the world. Moreover, we identified evidence for 3 distinct subtypes of SSA ESCC based on RNA expression, CNAs, and point mutation frequency. Finally, perhaps surprisingly given the important role of viral pathogens in the other most common cancers of SSA (e.g., Kaposi's sarcoma, cervical cancer, and Burkitt's lymphoma), we did not find evidence for a viral cause of SSA ESCC.

The finding of discrete ESCC subtypes based on RNA expression and CNA analysis was unexpected. This observation is not explained by available differences in tumor content, grade, or stromal contamination (Figure 1), and subtypes 1 and 2 were identified when the tumors were analyzed in 2 discrete cohorts prior to the aggregate analysis shown in Figure 4. Tumors of all the distinct subtypes clustered together when compared with other PanCan malignancies (Supplemental Figure 1) and appeared as similar classically squamous cancers by routine histopathological examination, although we think it is likely that at least subtypes 1 and 2 could be discerned using immunohistochemical markers. Subgroups 1a and 1b were distinguished by the frequency of CNAs, mutations per Mb sequenced, and frequency of *TP53* and *TP63* alterations (Figures 1 and 2). The subtypes were additionally distinguished by gene set analysis, with subgroup 1a showing a relative overexpression of DNA replication, repair, and recombination markers. In other studies, these transcripts are often associated with specific periods of the cell cycle (e.g., S-phase), suggesting that this subgroup may cycle more rapidly than other subtypes. Subgroup 1b demonstrated a relative underexpression of homeobox genes, and subtype 2 was characterized by increased expression of transcripts associated with neural differentiation (Figure 4). It is unclear if these ESCC subtypes are restricted to SSA, or are present in esophageal cancers of other regions.

Several lines of evidence suggest that an unrecognized, occult carcinogen accounts for the epidemic of ESCC in SSA. First, tobacco use does not explain ESCC in SSA. Prior epidemiological studies have demonstrated that tobacco use is not a likely cause of most ESCC in SSA (6), and other classic smoking-associated cancers (e.g., bladder, lung, non-HPV head and neck) are not common in SSA. In our cohort, the majority of patients did not use tobacco, either smoked or chewed, and we did not see evidence for the known mutational spectrum of inhaled tobacco in the somatic tumor exomes. Likewise, our data exclude an integrated or episomal virus as a cause of SSA ESCC. In contrast, we did observe a mutational signature in our sample that is similar to an unusual signature that has been previously reported in a handful of cases (signature 29, COSMIC) (29). To date, this signature has been observed in only 4 Western patients with oropharyngeal squamous cell carcinoma, but was present in a substantial fraction of Malawian patients. In aggregate, we believe this analysis suggests an occult environmental agent, which is not inhaled tobacco or HPV, that substantially contributes to ESCC in Malawi. Since there are 100,000s of cases of ESCC in SSA annually, and the population of this region is rapidly growing, this undetermined carcinogen is of considerable importance from a global public health perspective.

Effectively treating or palliating the extremely high burden of ESCC poses a major challenge to health systems throughout SSA, especially given limited radiotherapy, endoscopy, and thoracic surgery capabilities. Moreover, prevention efforts are considerably hampered by poor understanding of why the incidence of ESCC is so high in the region. Applying modern molecular approaches to ESCC and other cancers can complement classical epidemiologic studies, by helping support or disprove possible etiologic contributors, and ultimately informing population-level strategies to prevent this almost uniformly fatal cancer in SSA when it occurs. Therapeutic targets might also be suggested in a region where poor supportive care

environments limit the utility of conventional radiotherapy and cytotoxic chemotherapy to treat cancer. As in our work reported here, it is vital that the global oncology community continue to partner with SSA colleagues to apply modern molecular techniques to cancer research in the region, in a manner that benefits local individuals and populations.

Methods

Clinical/collection. From January 2011 to February 2013, informed consent was obtained from patients with suspected ESCC undergoing diagnostic upper endoscopy at 2 national teaching hospitals in Lilongwe or Blantyre. A study questionnaire and medical record review was conducted to obtain demographic details and risk factor data. Endoscopic biopsies obtained using forceps were divided into 2 formalin-fixed paraffin-embedded specimens for local histologic interpretation, and 1 research specimen that was frozen at –80°C. Paired whole blood (60 ml) was collected in EDTA and centrifuged for buffy coat isolation and freezing at –80°C.

Whole-exome sequencing. Genomic DNA was extracted by using a QIAGEN DNA/RNA kit and then sheared to a peak target size of 300 bp using a Covaris Focused-ultrasonicator. The DNA library was prepared and captured using the SureSelect XT Human All Exon V5 according to the manufacturer's protocols, and then sequenced by an Illumina HiSeq2000 with 100-bp paired-end reads.

DNA reads were aligned to the human hg19 genome assembly using BWA (37). After mapping, marking duplicates, local realignment around indels, and base quality score recalibration were performed by using the GATK tool kit (38). Combined analysis by using UNCeqR (22) and MuTect (39) was done to detect somatic nucleotide variations. Circular binary segmentation (CBS) and GISTIC (40) were used for somatic CNA analysis. Gene mutational signatures analysis was performed using the WTSI Mutational Signature Framework (26, 27), and the Chinese ESCC data set (15) was included and analyzed for comparison.

RNA sequencing. RNA was extracted and rRNA was removed by using the Ribo-Zero kit (Illumina). RNA libraries were prepared by using the Illumina TruSeq RNA Sample Preparation Kit v2 and then sequenced by the Illumina HiSeq2000. Reads were subjected to quality control as previously described (23). RNA reads were aligned to the human hg19 genome assembly using Mapsplice (41). Gene expression was quantified using RSEM (42) and within-sample normalized to a fixed upper quartile of total reads. All data were log2 transformed, median centered, and clustered using ConsensusClusterPlus (43). The optimal number of RNA expression subgroups was initially determined using silhouette width and principal component analysis. The classification to nearest centroids (ClaNC) (44) prediction method was used to robustly identify smaller gene sets differentially expressed between groups. Further gene set analysis was then performed using significance analysis of microarrays for sequencing (SAMSeq) (45). Heat maps were generated using Java TreeView (46).

Viral sequence detection. To detect viral sequences, RNA reads were first aligned to the human hg19 genome assembly. Unaligned sequences were processed by aligning to all viral sequences from the NCBI viral genome database and assembled by using Triniby (47). To search for possible unknown HPV strains in RNA reads, additional steps of aligning to all known HPV strains with relaxed mapping criteria were performed. All the sequences that share 70% similarity with any known HPV strains were manually checked against other possible alignments.

Data availability. To comply with restrictions on sequencing data put in place by the Malawi National Health Sciences Research Committee, the genomic data have been securely stored at the University of North Carolina. Data will be made available upon request to qualified investigators with approval from the Malawi National Health Sciences Research Committee.

Statistics. Statistical analyses were performed in each bioinformatic method as described by the relevant citation. RNA transcripts were log2 transformed and median centered before clustering. Differences among clinical and environmental characteristics, single-nucleotide polymorphisms, and CNAs between subtypes were calculated by ANOVA. A P value of less than 0.05 was considered significant.

Study approval. Patients over 18 years of age and willing to participate in this study provided written informed consent in their native language. Human studies approval was given by the Malawi National Health Sciences Research Committee, Lilongwe, Malawi and the University of North Carolina Internal Review Board, Chapel Hill, North Carolina.

Author contributions

NES, WL, JMS, CGS, and SG designed the research. YBM, SG, GM, NGL, LLW, and CGS executed studies in Malawi including patient identification and consent and specimen acquisition. NP and NGL performed histopathological evaluation of samples. WL, JMS, WRJ, KAH, MDW, and JSP analyzed data. NES, WL, SG, and JMS wrote the manuscript. All authors critically reviewed the manuscript.

Acknowledgments

We would like to thank Neil Hayes for comments on the manuscript. This work was supported by the UNC bioinformatics core. This work was supported by grants from the NIH (U54CA19015, P30CA016086, R01CA163896). Lindsey Wolf was supported by a Fellowship from the Doris Duke Charitable Foundation of the University of North Carolina. Carol Shores obtained funding from the North Carolina Translational & Clinical Sciences Institute through the NIH Clinical and Translational Science Awards at the University of North Carolina—Chapel Hill. Yohannie Mlombe was supported by the Fogarty International Clinical Research Scholars & Fellows Program of Vanderbilt University.

Address correspondence to: Norman E. Sharpless, The Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, CB #7295, Chapel Hill, North Carolina 27599, USA. Phone: 919.966.1185; E-mail: nes@med.unc.edu.

1. Ferlay J, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359–E386.
2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin*. 2015;65(2):87–108.
3. Msyamboza KP, et al. Burden of cancer in Malawi; common types, incidence and trends: national population-based cancer registry. *BMC Res Notes*. 2012;5:149.
4. Wolf LL, Ibrahim R, Miao C, Muyco A, Hosseinipour MC, Shores C. Esophagogastroduodenoscopy in a public referral hospital in Lilongwe, Malawi: spectrum of disease and associated risk factors. *World J Surg*. 2012;36(5):1074–1082.
5. White RE, Parker RK, Fitzwater JW, Kasepoi Z, Topazian M. Stents as sole therapy for oesophageal cancer: a prospective analysis of outcomes after placement. *Lancet Oncol*. 2009;10(3):240–246.
6. Mlombe YB, et al. Environmental risk factors for oesophageal cancer in Malawi: A case-control study. *Malawi Med J*. 2015;27(3):88–92.
7. Fagundes RB, et al. Higher urine 1-hydroxy pyrene glucuronide (1-OHPG) is associated with tobacco smoke exposure and drinking maté in healthy subjects from Rio Grande do Sul, Brazil. *BMC Cancer*. 2006;6:139.
8. Lubin JH, et al. Maté drinking and esophageal squamous cell carcinoma in South America: pooled results from two large multicenter case-control studies. *Cancer Epidemiol Biomarkers Prev*. 2014;23(1):107–116.
9. Kamangar F, Schantz MM, Abnet CC, Fagundes RB, Dawsey SM. High levels of carcinogenic polycyclic aromatic hydrocarbons in mate drinks. *Cancer Epidemiol Biomarkers Prev*. 2008;17(5):1262–1268.
10. Islami F, et al. Tea drinking habits and oesophageal cancer in a high risk area in northern Iran: population based case-control study. *BMJ*. 2009;338:b929.
11. Mark SD, et al. Prospective study of serum selenium levels and incident esophageal and gastric cancers. *J Natl Cancer Inst*. 2000;92(21):1753–1763.
12. Cook MB, Chow WH, Devesa SS. Oesophageal cancer incidence in the United States by race, sex, and histologic type, 1977-2005. *Br J Cancer*. 2009;101(5):855–859.
13. Esophageal cancer: epidemiology, pathogenesis prevention. *Nat Clin Pract Gastroenterol Hepatol*. 2008;5(9):517–526.
14. Gao YB, et al. Genetic landscape of esophageal squamous cell carcinoma. *Nat Genet*. 2014;46(10):1097–1102.
15. Lin DC, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat Genet*. 2014;46(5):467–473.
16. Song Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer. *Nature*. 2014;509(7498):91–95.
17. Sawada G, et al. Genomic landscape of esophageal squamous cell carcinoma in a Japanese population. *Gastroenterology*. 2016;150(5):1171–1182.
18. Hoadley KA, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*. 2014;158(4):929–944.
19. United Nations Statistics Division. 2015. http://data.un.org/CountryProfile.aspx?crName=malawi.
20. Msyamboza KP, et al. The burden of selected chronic non-communicable diseases and their risk factors in Malawi: nationwide STEPS survey. *PLoS One*. 2011;6(5):e20316.
21. United Nations Malawi AIDS response progress report. 2014. http://www.unaids.org/en/regionscountries/countries/malawi.
22. Wilkerson MD, et al. Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res*. 2014;42(13):e107.
23. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576–582.
24. Shiiba M, et al. Down-regulated expression of SERPIN genes located on chromosome 18q21 in oral squamous cell carcinomas. *Oncol Rep*. 2010;24(1):241–249.

25. Snijders AM, et al. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*. 2005;24(26):4232–4242.
26. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3(1):246–259.
27. Alexandrov LB, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.
28. Roberts SA, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet*. 2013;45(9):970–976.
29. Forbes SA, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43(Database issue):D805–D811.
30. Li X, et al. Systematic review with meta-analysis: the association between human papillomavirus infection and oesophageal cancer. *Aliment Pharmacol Ther*. 2014;39(3):270–281.
31. Hardefeldt HA, Cox MR, Eslick GD. Association between human papillomavirus (HPV) and oesophageal squamous cell carcinoma: a meta-analysis. *Epidemiol Infect*. 2014;142(6):1119–1137.
32. Yong F, Xudong N, Lijie T. Human papillomavirus types 16 and 18 in esophagus squamous cell carcinoma: a meta-analysis. *Ann Epidemiol*. 2013;23(11):726–734.
33. Syrjänen KJ. HPV infections and oesophageal cancer. *J Clin Pathol*. 2002;55(10):721–728.
34. Initiative for Vaccine Research (IVR): Viral Cancers: Epstein-Barr Virus. World Health Organization. http://apps.who.int/vaccine_research/diseases/viral_cancers/en/index1.html. Accessed September 14, 2016.
35. Hjalgrim H, Friborg J, Melbye M. The epidemiology of EBV and its association with malignant disease. *Human Herpesvirus: Biology, Therapy, and Immunoprophylaxis*. Cambridge: Cambridge University Press; 2007. http://www.ncbi.nlm.nih.gov/books/NBK47424.
36. Biggar RJ, Henle W, Fleisher G, Böcker J, Lennette ET, Henle G. Primary Epstein-Barr virus infections in African infants. I. Decline of maternal antibodies and time of infection. *Int J Cancer*. 1978;22(3):239–243.
37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
38. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–1303.
39. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213–219.
40. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12(4):R41.
41. Wang K, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38(18):e178.
42. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
43. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26(12):1572–1573.
44. Dabney AR. ClaNC: point-and-click software for classifying microarrays to nearest centroids. *Bioinformatics*. 2006;22(1):122–123.
45. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116–5121.
46. Saldanha AJ. Java Treeview--extensible visualization of microarray data. *Bioinformatics*. 2004;20(17):3246–3248.
47. Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–652.