scientific reports

Check for updates

OPEN

Exome sequencing identifies novel somatic variants in African American esophageal squamous cell carcinoma

Hayriye Verda Erkizan[1]✉, Shrey Sukhadia[2], Thanemozhi G. Natarajan[3], Gustavo Marino[4], Vicente Notario[5], Jack H. Lichy[6] & Robert G. Wadleigh[1,7]

Esophageal cancer has a strikingly low survival rate mainly due to the lack of diagnostic markers for early detection and effective therapies. In the U.S., 75% of individuals diagnosed with esophageal squamous cell carcinoma (ESCC) are of African descent. African American ESCC (AA ESCC) is particularly aggressive, and its biological underpinnings remain poorly understood. We sought to identify the genomic abnormalities by conducting whole exome sequencing of 10 pairs of matched AA esophageal squamous cell tumor and control tissues. Genomic analysis revealed diverse somatic mutations, copy number alterations (SCNAs), and potential cancer driver genes. Exome variants created two subgroups carrying either a high or low tumor mutation burden. Somatic mutational analysis based on the Catalog of Somatic Mutations in Cancer (COSMIC) detected SBS16 as the prominent signature in the high mutation rate group suggesting increased DNA damage. SBS26 was also detected, suggesting possible defects in mismatch repair and microsatellite instability. We found SCNAs in multiple chromosome segments, encoding *MYC* on 8q24.21, *PIK3CA* and *SOX2* on 3q26, *CCND1*, *SHANK2*, *CTTN* on 11q13.3, and *KRAS* on 12p12. Amplifications of *EGFRvIII* and *EGFRvIVα* mutants were observed in two patients, representing a novel finding in ESCC that has potential clinical relevance. This present exome sequencing, which to our knowledge, represents the first comprehensive exome analysis exclusively in AA ESCC, and highlights novel mutated loci that might explain the aggressive nature of AA ESCC and lead to the development of diagnostic and prognostic markers as well as therapeutic targets.

Esophageal cancer (EC) is one of the most lethal cancers, primarily due to the lack of early diagnostic markers and effective treatments. EC represents 26.3% of aerodigestive malignancies in the U.S., thus imposing a substantial burden on a sizeable proportion of the population and the health care system[1]. In 2021, the ratio of expected deaths (15,530 cases) to expected new diagnoses (19,260 cases) is estimated to be 80%, and men are 3.5 times more likely to be diagnosed with EC than women[1]. EC is the seventh leading cause of cancer deaths in males, accounting for 4% of all cancer deaths[1]. For all stages, combined survival of esophageal cancer is comparable to liver cancer (20%) and slightly higher than pancreas cancer (10%), which has one the lowest survival rates among diverse cancers[1].

EC consists of two major histological subtypes, esophageal adenocarcinoma (EAC) and ESCC; each subtype predominantly affects a specific population. AA comprise about 75% of ESCC patients[2,3]. The 5-year survival rate for localized EC in AA patients has been found to be 25%, in contrast the survival rate for EC patients of Caucasian origin is 48%[4]. However, for metastatic disease, the survival rate decreases dramatically to 5% regardless of race[2]. The lack of access to quality health care and low socio-economic status have been reported to partly contribute to the higher incidence and mortality rates of ESCC among AA[5,6].

[1]Institute for Clinical Research, Veterans Affairs Medical Center, Washington, DC, USA. [2]Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA. [3]Queromatics, Laveen, AZ, USA. [4]Hepatology and Gastroenterology, Veterans Affairs Medical Center, Washington, DC, USA. [5]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA. [6]Pathology and Laboratory Service, Veterans Affairs Medical Center, Washington, DC, USA. [7]Hematology and Medical Oncology, Veterans Affairs Medical Center, Washington, DC, USA. ✉email: hayriye.erkizan@va.gov

EC is caused by the combined effects of genetic and environmental risk factors, including tobacco use, alcohol consumption, and certain dietary habits[7]. Various EC studies have disclosed gender, racial, socioeconomic, regional disparities, and variability in worldwide geographical incidence rates[7,8]. ESCC is endemic in parts of Asia such as China, India, Iran, South Africa, and South American countries[9].

Various studies revealed molecular evidence for the heterogeneous nature of ESCC[10], yielding three main molecular subgroups[11]. Group 1 (ESCC1), or the "classical" subtype, closely resembles head and neck and lung squamous cell carcinomas[11], and has been shown to be the most common in Asian patients. ESCC1 is typically associated with mutations in the genes involved in oxidative stress and detoxification pathways, and the amplification of *SOX2* and *TP63*. Group 2 (ESCC2), occurs primarily in Eastern European and South American patients that carry frequent mutations in *NOTCH1*, *ZNF750*, *KDM6A*, and *KMT2D* [11]. Group 3 (ESCC3) has been found in very few patients, for example, one study reported five AA patients that displayed, activation of PI3K pathway, and *SMARCA4* mutations and relatively low occurrence of *TP53* mutations[11]. These subtypes highlight dis-similarities that exist between Asian, Caucasian, and AA ESCC[12–14].

Defects in molecular and genetic mechanisms associated with AA ESCC are not well defined due, in part, to its under-representation in epidemiological studies[11,12,15]. Hence, the biological basis for the lethality and aggressiveness of AA ESCC remains to be fully understood[16]. In our earlier studies, we conducted the first comparative genomic hybridization (CGH) analysis in AA ESCC that revealed widespread chromosomal imbalances and prominent abnormalities throughout the AA ESCC genome[17]. We performed gene expression profiling in AA ESCC tumors, which revealed a profound disruption of genes involved in various pathways including stress response and detox pathway, integrin signaling, and protein ubiquitination[18]. This study identified uniquely impaired biological processes in AA ESCC, which partially overlapped with findings in ESCC patients of Asian origin[18].

In our current study, we sought to identify somatic mutations in the AA ESCC genome by whole-exome sequencing (WES). We identified two subgroups of ESCC based on tumor mutation burden (TMB) and revealed recurrent novel SCNAs in cancer-related genes. We performed analysis based on COSMIC and detected signatures SBS16 and SBS26 that might contribute to AA ESCC pathogenesis. These deleterious alterations may play a role in the aggressive nature of the disease in the AA population. Further analysis in a larger set of AA ESCC may lead to the identification of AA ESCC specific clinical biomarkers and therapeutic targets.

Result

Whole-exome sequencing reveals a complex mutation profile, and SCNAs in AA ESCC. Whole-exome sequencing was performed on matched normal-tumor samples from 10 AA patients (nine males and one female) with advanced-stage ESCC, with ages ranging from 53 to 80 years (Supplementary Table S1). All patients, except for one, reported tobacco use and alcohol consumption. We employed an analysis pipeline delineated in Supplementary Fig. S1 that involved rigorous quality control (QC) and filtering mechanisms, along with methods described in the Genome Analysis Toolkit (GATK) Best Practices (https://gatk.broadinstitute.org/hc/en-us). Somatic single nucleotide variants (SNVs) and short insertion-deletions (InDels) called by at least two of three variant calling algorithms were filtered by a read-depth of 50× or higher. This analysis included variant allele frequency (VAF) > 5% and rare variants with < 1% minor allele frequency (MAF) in African population (Supplementary Fig. S1).

We identified predominantly C > T substitution in the majority of the samples (Fig. 1a). Transition to transversion ratio (Ti/Tv) was 2.4 (Fig. 1b). Samples T9 and T15 demonstrated significant C > A transversion (Fig. 1c). Missense mutations (N = 3682) constituted the primary type of alteration in the coding region of tumor samples (Fig. 1d). Nonstop (stop-loss) was rare (N = 23). A median of 270 variants per sample was revealed in our cohort (Fig. 1e) (Supplementary Table S2). Five samples (T1, T6, T5, T7, and T14) displayed a high tumor mutation burden (TMB) referred to as High Mutation Rate (HMR) samples. In each of these samples, nonsynonymous mutations ranged from 377 to 813 (Table 1). In contrast, low TMB or Low Mutation Rate (LMR) was detected in T8, T9, T15, T16, and T19, each sample showed a median of 29 (ranging from 5 to 44) nonsynonymous coding mutations (Table 1).

Comparison of the mutation rate of the AA ESCC cohort with the TMB of diverse types of cancer in the TCGA database showed that the median mutation rate of AA ESCC was greater than most tumors, which ranked between skin cutaneous melanoma and lung squamous cell carcinoma (Fig. 1f).

We performed somatic mutational signature analysis[19] based on COSMIC and identified SBS16 in all our AA ESCC study samples, thus representing the predominant mutational signature (Fig. 2a, b, yellow portion of the bar graphs). Eight tumor samples displayed the highest contribution to SBS16 (Fig. 2a, b). SBS16 may be generated by inefficient nucleotide excision repair and elevated levels of DNA damage suggesting the involvement of these mechanisms in AA ESCC[19]. In addition to S16, Signatures 1, 5, and 26 were observed in the HMR samples (Fig. 2a). SBS1 suggests failure to repair the product of spontaneous or enzymatic deamination of 5-methylcytosine to thymine. Consequently, S1 represents a clock-like feature for cancer tissues, termed mitotic clock, and may be correlated with the age of the individual[20]. SBS1 is usually co-observed with SBS5, as seen in our HMR cases. SBS5 is another clock-like signature that correlates with the individual's age but not with the mitotic clock. The etiologic factors causing SBS5 are uncertain, but the effect of tobacco smoking is suspected[19], interestingly, nine of the 10 cases in our current study were tobacco smokers. We observed 10% contribution of SBS26 to the mutational profile of HMR samples (Fig. 2a, light blue). Defective DNA mismatch repair and microsatellite instability (MSI) contribute to SBS26[21].

The presence of SBS26 in the HMR group led us to investigate microsatellite instability (MSI) using MSI sensor on paired normal and tumor sequence data[22]. This analysis demonstrated high MSI scores (%15) in

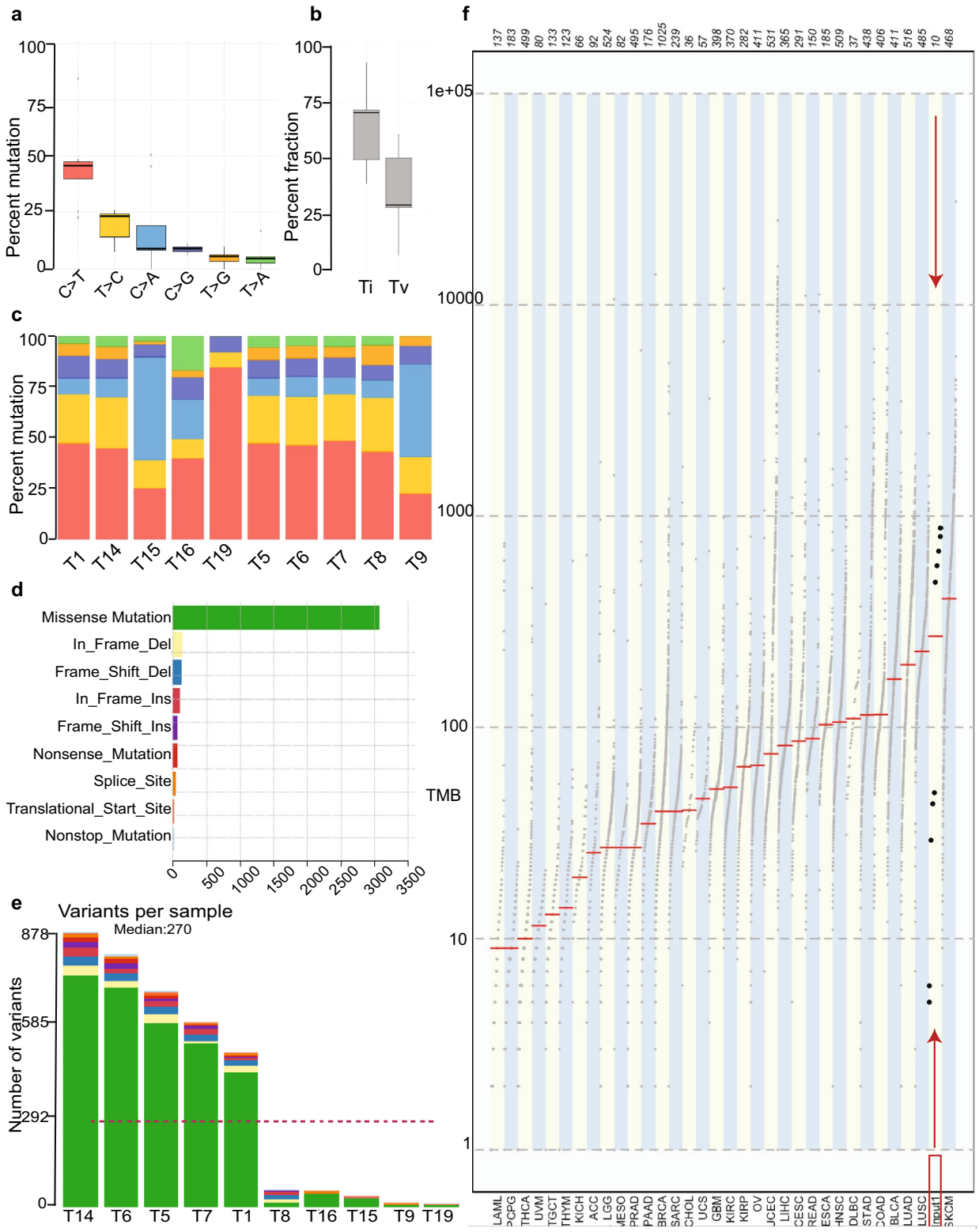

Figure 1. Summary of exome sequencing and tumor mutation burden (TMB). (**a**) Proportion of somatic substitution types in processed and filtered SNVs. (**b**) The ratio of somatic transition to transversion of AA ESCC cohort. (**c**) The proportion of somatic substitution types (C > T, red, T > C, yellow, C > A, blue, C > G, purple, T > G, orange, T > A, green) in each tumor samples. (**d**) The proportion of mutation types in the samples, missense mutations (green), in frame deletions (yellow), frameshift deletions (blue), in frame insertions (brick red), frameshift insertions (purple), nonsense mutations (bright red), changes in the splice sites (orange), and nonstop mutations (light blue). (**e**) Number and types of variants found in each sample. (**f**) The comparison of mutation burden of AA ESCC pilot cohort across TCGA datasets. Our samples are indicated by red rectangle and arrows. Y-axis is the total number of non synonymous mutations found in each tumor sample.

Tumor	2 or more variant caller (>50DP, 5>VAF)		
	#SNV < 1%AFR	# of InDel	Total coding mutations
T14	3460	1940	878
T6	3339	1239	804
T5	2716	1243	686
T1	2152	1033	588
T7	1663	956	491
T16	125	9	49
T15	80	1	44
T8	93	629	29
T9	22	0	6
T19	13	0	5

Table 1. Number of mutations* in each tumor sample. *Mutations called by two or more algorithms, filtered for rare (< 1% MAF) African population frequency, and sorted by somatic nonsynonymous coding mutations.

tumor samples with SBS26 signature (Table 2). The presence of SBS26 was highly correlated with the MSI scores ($r^2 = 0.9908$) (Fig. 2c).

In LMR samples, we found other mutational signatures such as SBS18, possibly due to DNA damage by reactive oxygen species in two tumor samples (T9 and T15), SBS4, due to tobacco smoke carcinogens in three tumor samples, and SBS8, possibly due to defective homologous recombination repair in a tumor sample from a female ESCC patient (Fig. 2b).

SCNAs in AA ESCC genomic regions encode cancer-related genes. Our prior CGH analysis on 17 AA ESCC tumors revealed a preponderance of gains and losses in multiple chromosomal regions suggesting the involvement of a plethora of SCNAs[17]. In the present study, we evaluated SCNAs in an orthogonal panel of tumor samples from mostly late-stage AA patients with ESCC. Multiple regions of copy number gain and loss were observed in seven samples: T1, T5, T6, T9, T14, T16, and T19 (Fig. 3a and Supplementary Fig. S2). These SCNA-rich samples displayed on an average of 62 (ranging from 46 to 90) different chromosomal regions. Tumor samples (T1, T7, T8, and T15) with fewer copy number aberrations showed on average 15 SCNAs (ranging from 5 to 25), presenting fewer copy number aberrations. Sample T5 and T9 had a moderate number of SCNAs (Fig. 3a).

The most recurrent copy number changes along the AA ESCC genome were seen in the amplification of chromosomes 3q, 8q, 11q, and 12p (Fig. 3b). Three tumor samples (T5, T16, and T19) exhibited high-level amplification in 3q26, including *WWTR1*, *TP63*, *PIK3CA*, *SOX2*, *SOX2-OT*, *and ZNF639* genes (Fig. 3b and Supplementary Fig. S3a, b). We also observed an extended 3q amplicon, a distinct ESCC1 feature, in 30% AA ESCC patients. The region 8q harboring the proto-oncogene *MYC* was amplified in T5, T14, and T16 (Fig. 3b and Supplementary Fig. S3b). One of the highly amplified regions in chromosome 11q13 encoding several critical genes, including *CCND1*, *FGF3*, *FGF4*, was observed in T5, T6, T14, T16, and T19 (Fig. 3b and Supplemental Fig. S3a, b). Samples T14, T19, and T6 revealed amplification on the short arm of chromosome 12, including 12p12.1 that harbors *KRAS*, *SOX5*, *ARNTL2*, *BCAT1* (Supplementary Fig. S3b).

T16 displayed the highest copy number amplification with eight copies of the 2q33.1 region encoding *CASP10* and *CFLAR*, both of which function in cell death[23]. The second-highest amplification number with seven copies of five different chromosome regions encode for genes including *SOX2*, *ANO1*, *FADD*, *CTTN*, *POLD3*, *RNF169*, *XRRA1*, *PAX9* in sample T19 in this study (Supplementary Table S3). A rare, distinctive homozygous deletion of the entire chromosome 2 was displayed by T1 and T5 samples (Supplementary Fig. S3a). T1 and T5 additionally displayed a deletion of 22q (Supplementary Fig. S3a). Chromosome 22q harbors several cancer genes, including histone acetyltransferase *EP300* (E1A Binding Protein 300), a known ESCC driver gene[14]. Two tumor samples, T14 and T19, displayed a loss of 4p13 that encodes *RHOH*, a member of the RAS superfamily, and *PHOX2B*, a homeobox transcription factor (Supplementary Fig. S3b). Our previous CGH analysis revealed the complete loss of chr4 in all 17 ESCC samples (Fig. 3b)[17]. Additionally, T19 carries a deletion of 18q21 harboring *SMAD4*, a region that we found deleted in another set of AA ESCC tumors[24]. Amplification on the short arm of chromosome 2 that contains a super-enhancer *ZFP36L2* gene was observed in three samples (T6, T14, and T19) and amplified in 30% of AA ESCC tumors in our previous aCGH study.

Chromosomal copy number gains in 1q distal, 2p proximal, 3q distal, 5p proximal, 8q distal, 11q, 16p, 17p and 17q, 18p, 19p were also found in 17 AA ESCC cohort[17]. Chromosomal copy number losses, also observed in 17 AA ESCC cohort, included 1p distal, 2qdistal, 3p distal and 3q proximal, whole chr4, whole 5q, 6q proximal, 9p distal, 11p proximal, 11qdistal, 13q, 14p proximal regions. The regions also found in 17 AA ESCC cohort indicated by arrows in Fig. 3b.

Although each patient's tumor sample revealed a unique set of genomic alterations, a clear pattern of shared regions of SCNA loci between tumors was evident in the tumor samples from the ten patients. The correlation analysis by Pearson's method demonstrated two hierarchical clusters of AA ESCC according to their SCNA profile (Fig. 3c). Copy number-rich samples T6, T14, T19, and T16 having more common SCNAs were clustered together and formed Cluster 1. Except for T16, Cluster 1 samples showed 12p amplification, a region that encodes *KRAS*. The second cluster, Cluster 2, contained samples with fewer copy number changes and is further divided

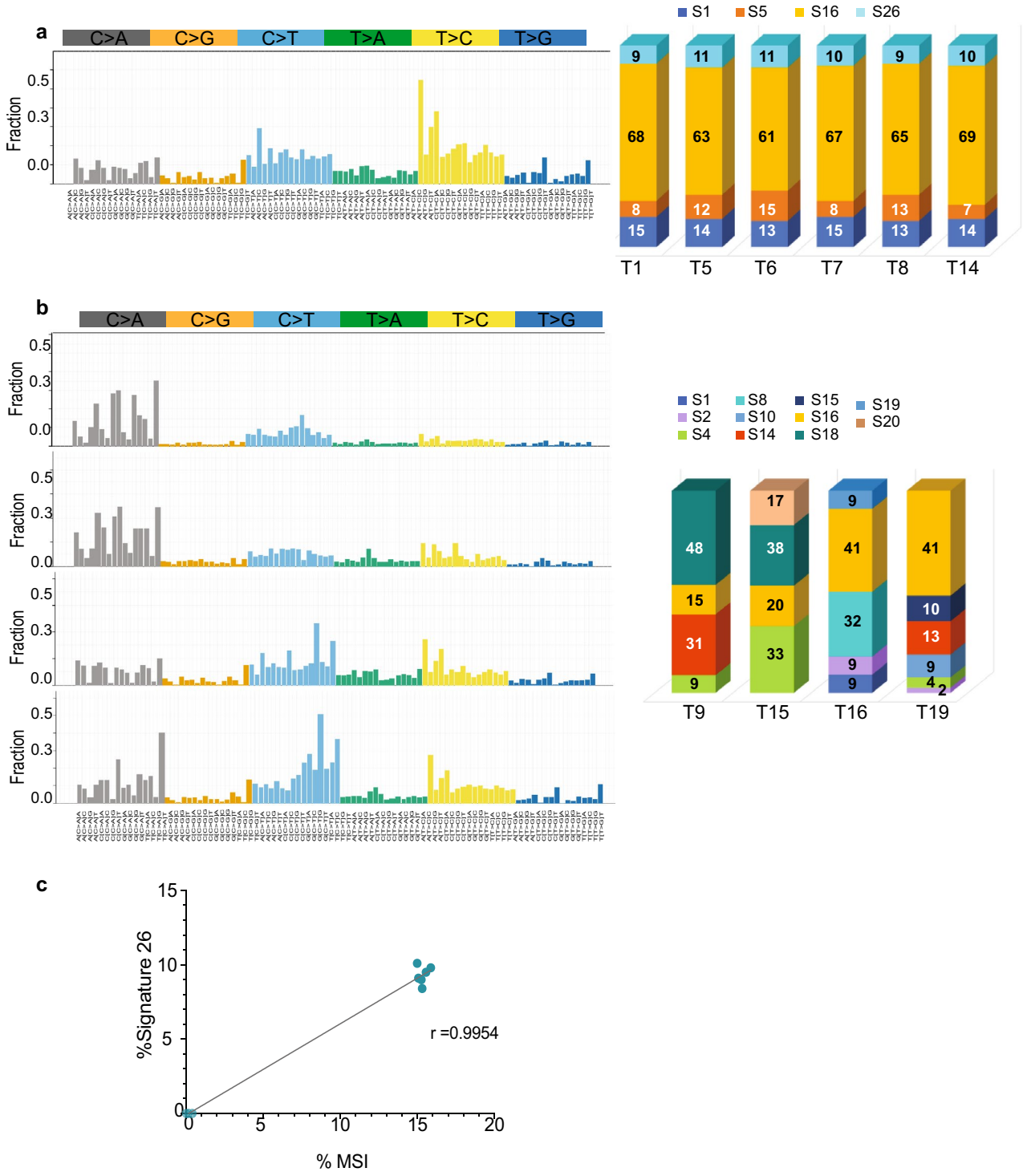

Figure 2. Mutational Signatures based on COSMIC Signatures extrapolate the nature of mutational processes in AA ESCC. (**a**) T > C mutations at ATN trinucleotides constitute Signature 16, which is demonstrated by yellow color in the bar graphs showing the fraction of mutational signatures contributed to whole mutations in each tumor sample. Signatures 1, 5, 16 and 26 were found in similar proportions (percentage) across the samples of T1, T5, T6, T7, T8, and T14. (**b**) Samples T9, T15, T16, and T19 contained various mutational signatures, each tumor sample had a unique mutational signature combination. Bar graphs shows the percentage of each signature in the tumor samples. (**c**) Signature 26 was correlated with percent microsatellite instability (%MSI) in Signature 16-high samples. Pearson correlation coefficient *r* was 0.9954 and 95% confidence interval 0.9799 to 0.9989. R^2 was calculated as 0.9908.

Tumor	% S26	%MSI
T1	9	15.3
T5	11	15.9
T6	11	15.0
T7	10	15.3
T8	10	15.1
T14	10	15.6
T9	0	0.4
T15	0	0.2
T16	0	0.1
T19	0	0.0

Table 2. Percent signature 26 and microsatellite instability in AA ESCC.

into subclusters. Cluster 2A contained tumor samples T7, T8, T9, and T15, which harbored mostly focal copy number changes. Cluster 2B was formed by samples T1 and T5 containing broader changes such as deletion of chromosomes 2 and 22.

Rare mutation and amplification in *EGFR* gene. Two tumor samples, T15 and T5, carried *EGFR* mutations and amplification. Sample T15 harbored *EGFRvIII* mutation caused by a deletion of exons 2–7, in addition to an amplification of the region, chr7:55087058-chr7:55223523 on which *EGFR* gene is located (Fig. 4a). Sample T5 was found to carry *EGFR vIVa* which represents deletion of exons 25–27, and amplification in chr7:55268106-chr7:55272949, which includes *EGFR* (Fig. 4b). In samples T5 and T15, we observed a mutual exclusive amplification pattern between *EGFR* and *KRAS* genes. However, amplification of cell cycle-related genes and *MYC* co-occurred with *EGFR* and with *KRAS* amplification.

Single nucleotide variations and short InDels. The high mutation burden in half of our AA tumor samples led to the detection of mutations in 6169 unique genes of which approximately 50% carried one mutation per gene. Analysis of frequently mutated genes showed that the top 45 genes were found in seven tumor samples (Fig. 5). The most frequently mutated gene in six samples was *ZDHHC11* (zinc finger DHHC-type containing 11) located at 5p15.33. Although the gene carried various types of mutations within the same sample, we suspected the passenger nature of these mutations in *ZDHHC11* partly due to scarce reports on mutated *ZDHHC11* in cancer and the potentially benign consequence of these mutations. The frequency-based identification of significant genes was not possible because of the small size of our cohort and the high mutation rate in some of the tumor samples.

To prioritize cancer-related genes in AA ESCC, we employed a combination of analyses in Opencravat[25]. We identified 23 missense and splice site mutations and 34 noncoding or synonymous variants previously described in ESCC in COSMIC database (Supplementary Table S4). The most notable ones were TP53 p.Ile232Asn and ZFP36L2 p.Ser105Leu, both of which may be damaging as predicted by Combined Annotation-Dependent Depletion (CADD)-Phred score of 31 and 24.6, respectively. Further analyses with multiple tools, such as CADD[26], and annotation of deleterious genetic variants using neural networks (DANN)[27] and maftools[28] revealed damaging mutations in protein coding regions. These variants were overrepresented in several genes involved in various pathways including WNT, NOTCH signaling, TP53 tumor suppressor pathway, apoptosis, cell to cell communication and cell motility (Fig. 6 and Supplementary Tables S5 and S6).

We then searched for significant mutually exclusive or co-occurring pairs of genes (Supplementary Fig. 4). We did not detect any significantly mutually exclusive gene pairs. Among 23 co-occurred pairs which we detected in our current AA ESCC tumor samples, mutant *MUC4* and *MUC12* pairing is interesting as this pair was previously observed in smoking-associated non-small cell lung cancer patients[29]. This finding may illustrate the mutagenic role of tobacco in AA ESCC.

Candidate cancer driver genes in AA ESCC. To predict candidate cancer driver genes and prioritize cancer-related genes in AA ESCC, we employed a combination of analysis algorithms that included CancerGenomeInterpreter (CGI), 20/20+, CHASMplus, and dNdScv. These tools predicted *TP53* (p.Ile232Asn), *NCOR1* (p.Leu2168Phe), *APC* (p.Ser338Tyr), *KMT2C* (p.Gly315Ser), *CDKN1B* (p.Met52Arg), *NOTCH1* (p.Arg353Cys) as possible drivers (Supplementary Table S7).

In addition to the DNA binding domain mutation (p.I232N) of *TP53*, a mutation in Chr17.g.7576852:C > T led to the splice site loss creating an intron inclusion between exons 9 and 10 in two samples (T5 and T16). In T5, a second *TP53* splicing mutation (chr17.g.7673534:C > T) which is SNP rs11575997, was previously reported in various cancers in the COSMIC database, including five esophageal cancers. This variation creates a splice donor site mutation at exon 10 and is predicted to have a damaging effect on the protein, suggesting a pathogenic impact (https://www.ncbi.nlm.nih.gov/clinvar/RCV000785504.1). This variation creates a splice donor site mutation at exon 10 and is predicted to have a damaging effect on the protein, suggesting pathogenic impact (https://www.ncbi.nlm.nih.gov/clinvar/RCV000785504.1).

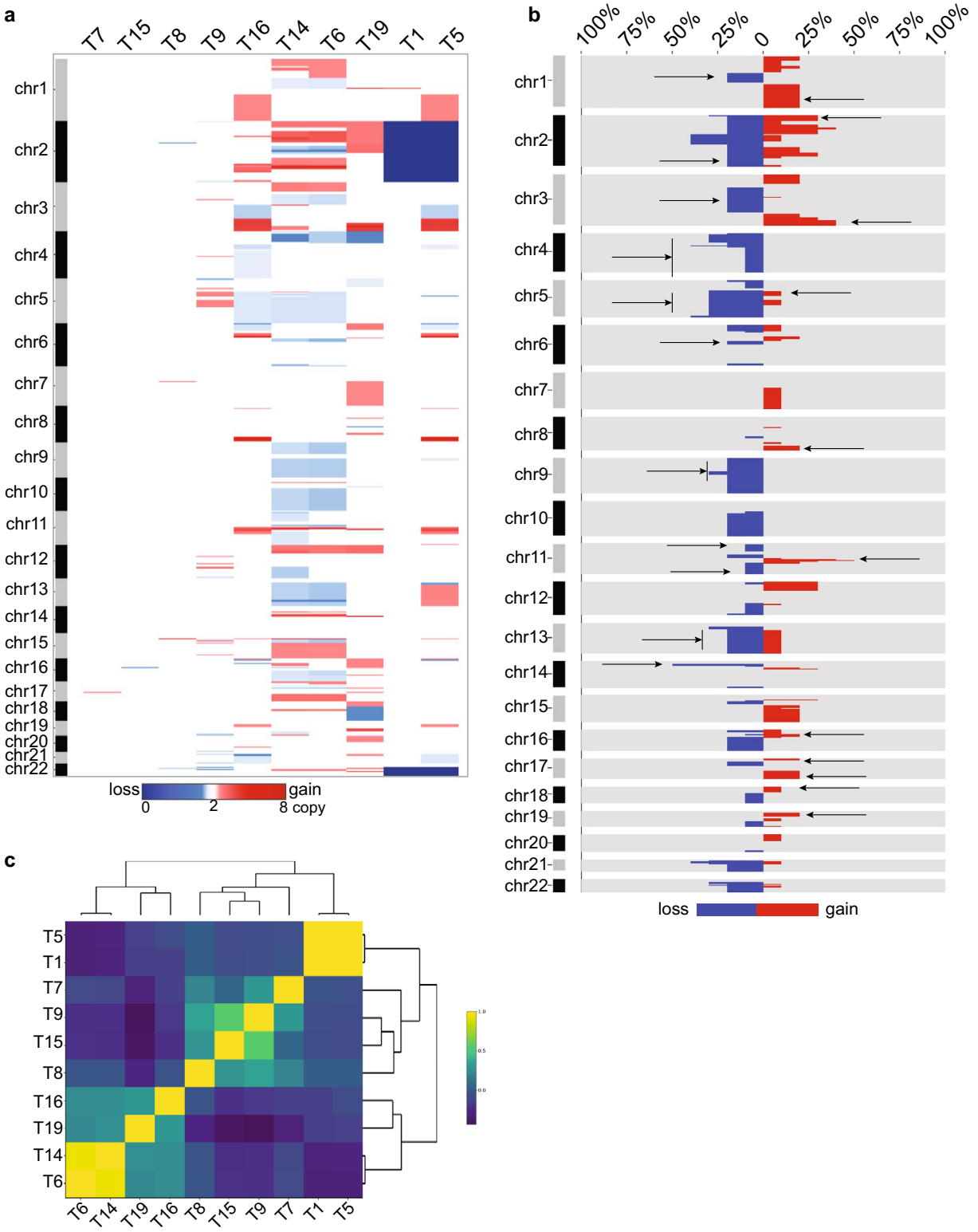

Figure 3. SCNAs in AA ESCC genomic regions encode cancer-related genes. (**a**) SCNA region profiles by 1 MB were visualized using CNApp. Dark blue represents homozygous deletion (copy number 0), white represents diploid (copy number 2), red represents copy number gains (the darkest red is copy number 8). (**b**) SCNA region frequency plot shows the percentage of samples with gains (red), losses (blue), and no alteration (gray) across chromosomes. Arrows represent the regions that also found in the 17 AA ESCC cohort analysis. (**c**) Plot shows the hierarchical clustering for samples correlation by Pearson correlation. Pearson *r* correlation coefficient $r = 1$ is yellow, $r = -1$ dark navy color.

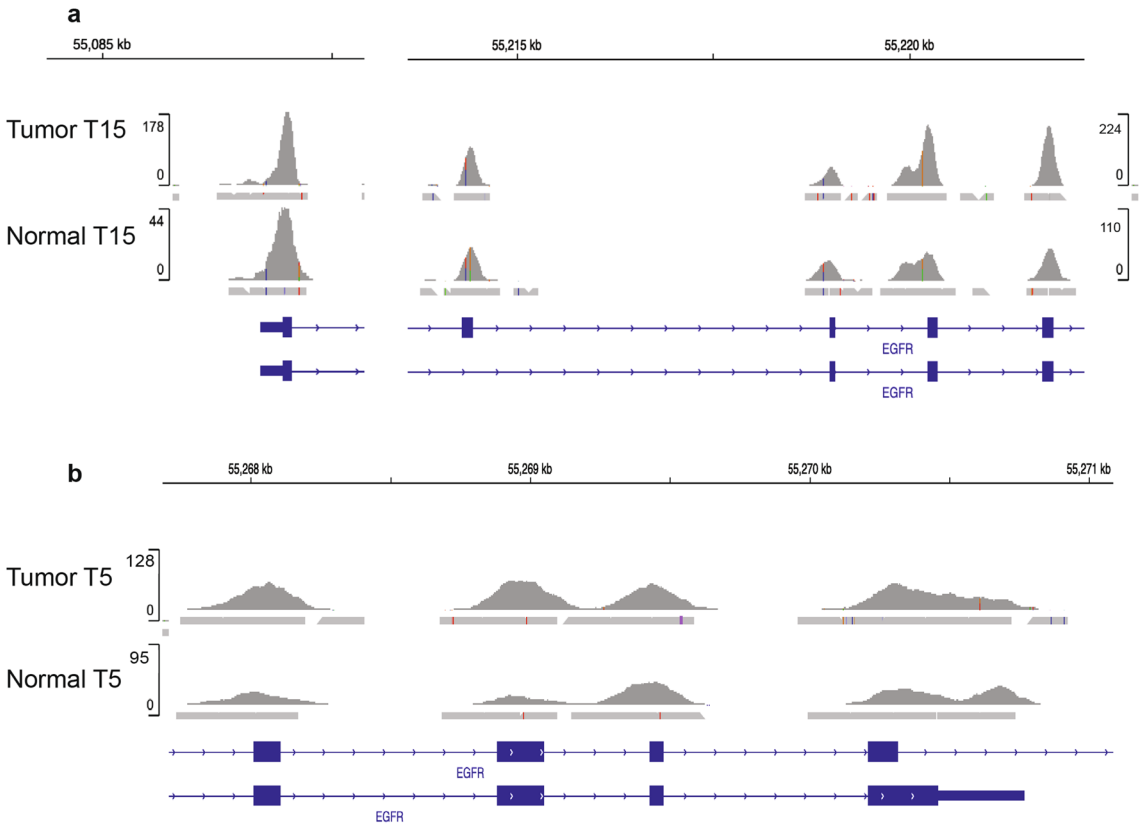

Figure 4. Amplified mutant EGFR in AA ESCC. (**a**) *EGFR* amplification in Sample T15 and corresponding normal sample sequencing readouts visualized in Integrative Genomics Viewer. (**b**) *EGFR* amplification in Sample T5 and corresponding normal sample sequencing readouts visualized in Integrative Genomics Viewer.

In summary, the current study identified two subgroups of ESCC based on tumor mutation burden (TMB), recurrent mutational signatures in HMR samples, and diverse signature profiles in LMR samples. We also detected MSI in HMR cancers. Our study revealed recurrent novel SCNA in cancer-related genes. SNVs and SCNAs suggested disruption of multiple pathways, including proliferation, cell cycle, epithelial to mesenchymal transition (EMT), invasion, and tumor metastasis in AA ESCC tumors (Fig. 6).

Discussion

ESCC in African Americans has high mortality rate. Despite the high morbidity and mortality associated with this disease, there is a paucity of genomic studies in AA ESCC. Previously, we laid the groundwork for investigating AA ESCC which revealed a high rate of genomic imbalances[17] and a dysregulated expression of genes in specific pathways[18]. The current study unmasked a multifaceted genomic landscape that defines a range of genomic alterations, and novel mutated loci that may potentially contribute to AA ESCC tumorigenesis.

We have conducted whole exome sequencing on 10 pairs of ESCC tumor and control tissues in AA, which to our knowledge, is the first comprehensive analysis of functional units in the ESCC genome exclusively in this ethnic group. Our results represent an important advance toward understanding the molecular basis for the lethality and aggressiveness of ESCC among African Americans.

AA ESCC displays a complex mutational profile that clustered half of our samples into a high mutation group, the other half formed a low mutation group. A similar distribution of high and low mutation rates has been reported in Sub-Saharan African ESCC patients[30]. Also, the former group exhibited a higher rate of mutation than the rates observed in Caucasian, Chinese, and Vietnamese ESCC patients[14,31]. Taken together, these findings are consistent with the high heterogeneity reported in ESCC tumors in patients of different ethnic origin.

An important novel finding in the current study is the identification of two EGFR mutations, *EGFRvIII* and *EGFRvIV*, accompanied by amplification in two AA ESCC tumor samples. While amplification of wild type *EGFR* has been previously reported in 7% of ESCC samples[32], *EGFR* mutations have been rarely reported in ESCC[33]. Both *EGFRvIII* and *EGFR vIVa* mutations lead to constitutive EGFR signaling, tumor growth and progression pathways[34–36]. These *EGFR* mutations along with amplification are characteristics of Glioblastoma multiforme[37] and wild type *EGFR* amplification was observed in ESCC with 6% frequency and associated with poor prognosis[38]. Consistent with previous reports, none of our AA ESCC tumor samples that displayed amplified mutant *EGFR* showed wild type *KRAS* amplification. The mutual exclusivity of *EGFR* amplification and *KRAS* mutations has been described in lung adenocarcinoma and colorectal carcinoma (CRC) and co-expression caused cytotoxic effect on cells[39,40]. Even though the incidence of *KRAS* amplification compared to its mutation is low, such as less than10% of patients with gastric cancer or CRC and 17% in EAC, clinical features of patients

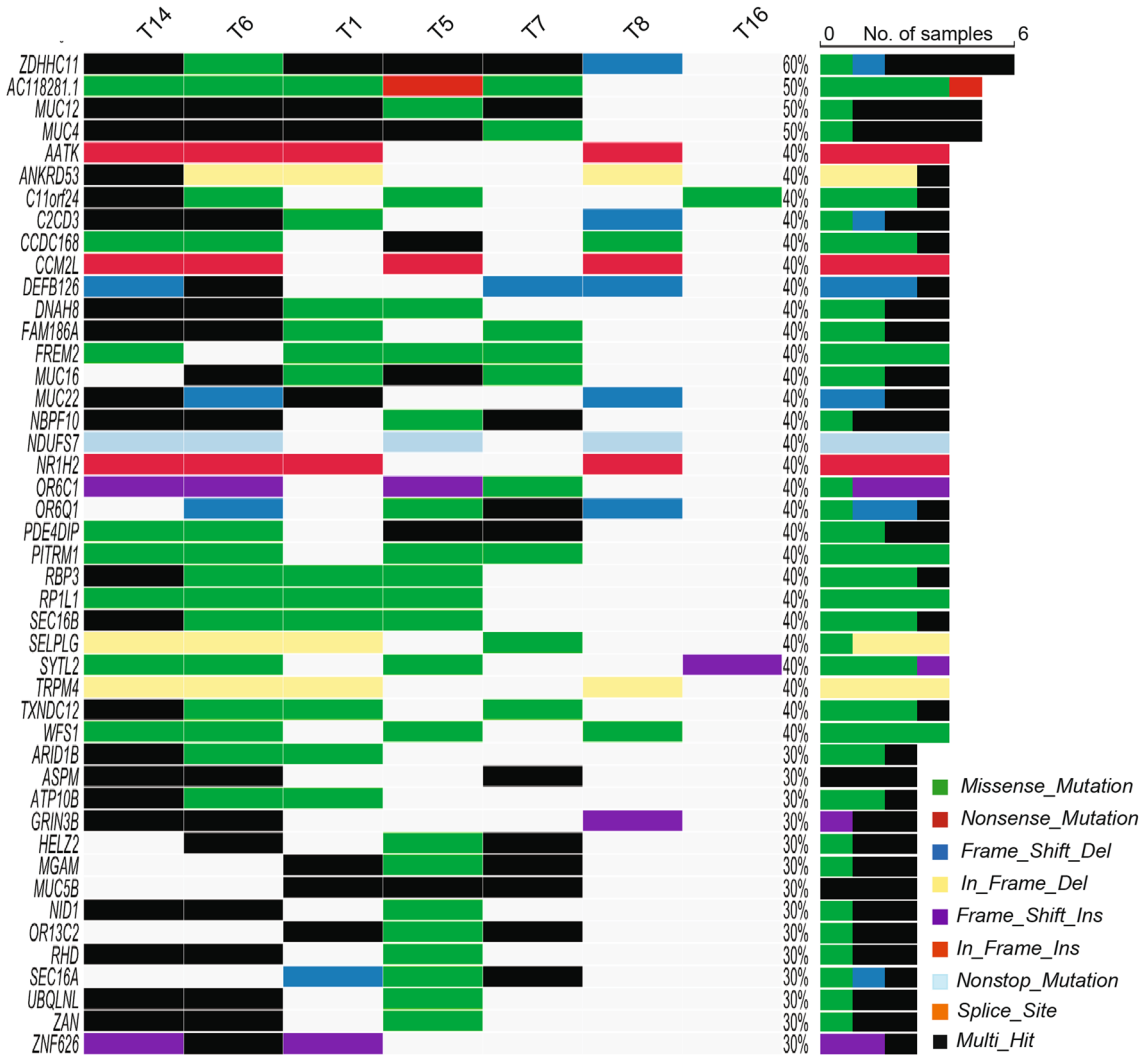

Figure 5. Single nucleotide variants of AA ESCC. Oncoplot by maftools was visualized mutations of missense, frameshift InDels, nonsense, splice site and translation start site that occurred in at least 30% of the tumors. Top 45 mutated genes were plotted.

with *KRAS* amplification are distinct, and the amplification is usually associated with poor prognosis in these patients[41–44]. The oncogenic effect of wild type *KRAS* amplification is mediated through the increased receptor tyrosine kinase-dependent activation of the Ras pathway in CRC[45]. A recent study in Asian ESCC shows a 6% wild type *KRAS* amplification and its mutual exclusivity from *EGFR* amplification[38]. Like in gastric cancer, EAC, and CRC patients, *KRAS* amplification is associated with worse survival in ESCC patients[38]. Similar to mutant *KRAS*, wild type *KRAS* amplification confers EGFR inhibitor resistance[41]. Therefore, ESCC patients may benefit from detailed *EGFR* profiling and determination of tumor *KRAS* amplification status before targeted therapies.

Amplification in 3q26 harboring *PIK3CA* and SOX2 was observed in 30% of the current AA ESCC cohort and reported in 40% AA ESCC by aCGH[17]. *PIK3CA* amplification is associated with poor prognosis in curatively resected patients with ESCC[46] and resistance to chemotherapy in other cancers[47]. *MYC* oncogene amplification was also observed in 30% of the AA ESCC tumor samples. MYC amplification is shown to be associated with lymph node metastasis and poor prognosis in ESCC[48]. *PIK3CA* and *MYC* amplifications co-occurred with either *EGFR* as observed in one sample or *KRAS* amplification in two samples in this study. The current study demonstrated 11q13 amplification, which is consistently amplified in 88% of AA and 68% of Asian patients[17,49,50]. Amplification of 11q13 is also revealed in a subset of HPV negative head and neck, lung, and esophageal squamous cell carcinoma[51] and breast cancer[52]. Recurrent amplification of 11q13 is associated with nodal metastasis in ESCC and poor prognosis in head and neck cancer patients[49,53]. Critical genes encoded by this region include *CCND1*, *FGF3*, *FGF4*, which are involved in cell-cycle regulation and tumor cell proliferation in oral squamous cell carcinoma, hepatocellular carcinoma, and ESCC[11,54–56].

Interestingly, 17p deletion, which is shared across all cancers, is rarely observed in African ancestry ESCC patients. Previously, we reported chromosome 17 gains in 17 AA with ESCC patients[17]. Another study on 51 South African ESCC cases did not display regional losses on chromosome 17[57]. However, other studies on Asian ESCC showed 17p deletions or losses up to 75% frequency[58,59], suggesting that chromosome 17 imbalance profile could be a distinguishing feature for ESCC between ethnic groups.

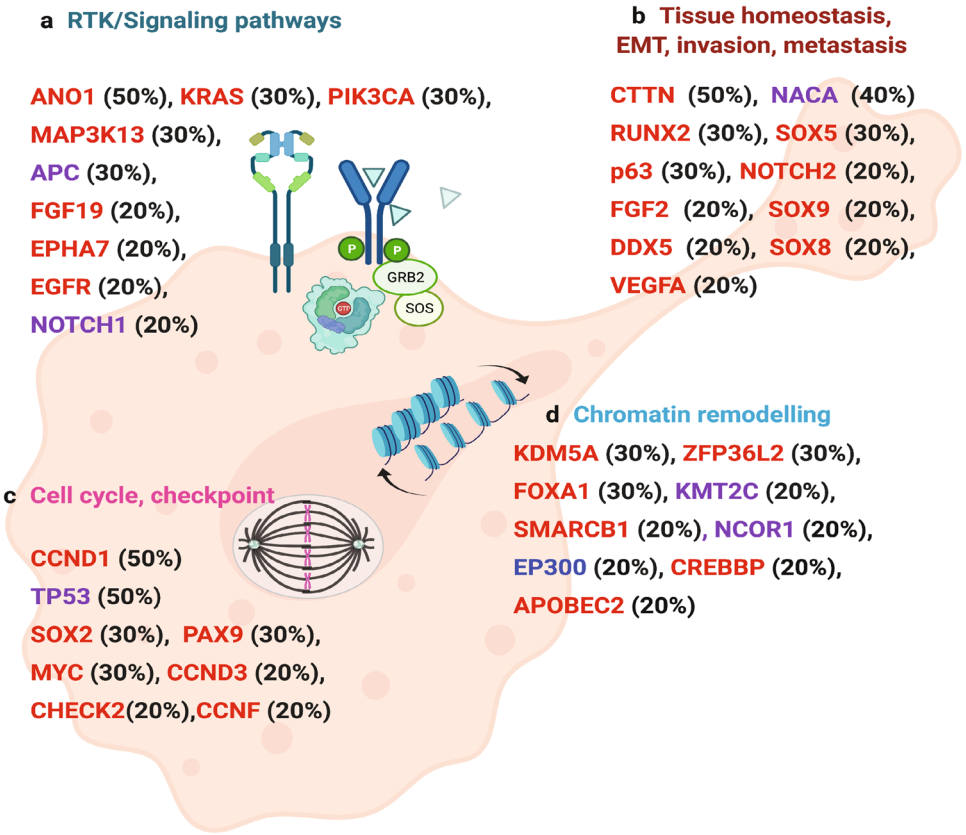

Figure 6. Summary of significantly altered genes in AA ESCC. (**a**) Genes that were mutated, or amplified in proliferation, cell cycle and cell cycle checkpoint controls. (**b**) Amplified genes found in squamous cell homeostasis, epithelial to mesenchymal transition (EMT), invasion and metastasis. (**c**) Genes that were amplified or mutated in receptor tyrosine kinase (RTK) and signaling pathways (**d**) Chromatin remodeling genes that were amplified, deleted, or mutated. Red designates amplification, blue indicates deletion, and purple shows missense mutations. Figure was created with BioRender.com.

In our dataset, homozygous deletion of the entire chromosome 2 and 22q in two AA ESCC patients was consistent with our previous CGH findings[17]. Chromosome 2 encodes 21 cancer hallmark genes[60]. The regions of deletion included several loci, including those coding for *LRP1B* (LDL receptor-related protein 1B) on 2q22.1-q22.2 and *NFE2L2* (nuclear factor, erythroid 2 like 2) on 2q31.2, both of which have been implicated in ESCC carcinogenesis[61,62]. Loss of chromosome 4p detected in the current study has been in AA ESCC samples[17]. The deletion of *SMAD4* in 18q21 is rarely found in ESCC in other ethnic datasets, albeit most frequently observed in EAC[63].

Ethnicity-based differences in mutation frequency in specific genes are observed in ESCC[14]. For example, *TP53*, *EP300*, and *NFE2L2* showed a significantly higher mutation frequency in Asian ESCC patients (ESCC1) than Caucasians (ESCC2)[14]. The most commonly mutated gene in ESCC is *TP53* (more than 70% of all samples) in all other cohorts[11]. In the current study, *TP53* was found to be mutated in 50% of the patient samples. However, none of these mutations were located in recognized ESCC hotspots.

Fractions of mutations found in each trinucleotide context constitute a mutational signature profile that may infer mutational processes and etiologic agents leading to tumorigenesis[19]. Signature SBS16, the most frequent mutation signature in our cohort, is significantly correlated with alcohol consumption in liver cancers, head and neck squamous cell carcinoma, and ESCC, leading to a high mutation burden[64–69]. However, the dominance of Signature 16 among other signatures in our dataset is surprising; in more than half of the samples, Signature 16 is the predominant signature, comprised 60% of the total mutational profile. This might suggest the strong mutagenic effect of alcohol in AA ESCC patients. Interestingly, we did not observe the APOBEC signature, which is more frequently seen in ESCC tumor samples[70,71]. Further studies should focus on the molecular mechanisms of alcohol consumption by illuminating the direct effect of alcohol and alcohol metabolites on ESCC especially in AAs, since this may explain the increased frequency of ESCC in AA population. SBS16 may represent defects in nucleotide excision repair and increased DNA damage. Our mutational analysis also identified SBS26, a mutational signature associated with DNA repair and microsatellite instability.

A limitation of our current study is the small sample size. Validating these findings in a larger sample collection is necessary and warranted. Another limitation is that, with one exception, all other samples were derived from advanced-stage tumors. Due to the small sample size and apparent high mutation rate, there is a possibility

of false-positive calls. Therefore, we set our variant call threshold to a read depth of 50x, which may lead to the under-representation of rare and highly heterozygous calls. Future studies in large sample size and single-cell studies may resolve heterogeneity issues. Epigenetics studies such as DNA methylation and chromatin accessibility might also shed light on the tumorigenesis and prognosis of ESCC in African Americans.

In summary, our present study, which to our knowledge, is the first comprehensive exome sequencing analysis exclusively directed at AA ESCC has revealed intriguing and novel somatic alterations and copy number variations. The associated mutational signatures have not been previously described in this particular cohort and are of importance, molecularly and clinically as they may, in part, form the basis for the aggressive nature of ESCC in African Americans. Unraveling key genetic changes that are pivotal to inception, progression, and metastasis could generate novel early diagnostic markers and actionable targets for efficacious intervention and drug discovery. Future studies are needed to illuminate and validate these findings that could eventually contribute to Precision Oncology in AA ESCC.

Materials and methods

Materials. Whole-exome sequencing of matched tumor- and normal-cell DNA from endoscopic biopsies or surgical specimens from ten AA ESCC patients were performed. The staging of the patients was done according to the methods described in the American Joint Committee on Cancer Staging Manual[72]. The District of Columbia Veterans Affairs Medical Center Institutional Review Board approved this study. Written informed consent and the approval for publication of results was obtained from each patient before the procedure. All experiments were performed in accordance with relevant guidelines and regulations.

DNA extraction and whole-exome sequencing. Genomic DNA extraction from frozen tissues was done using the MasterPure DNA extraction kit (Epicentre Technologies Corp., Madison, WI). To proceed to exon capture, we enriched the samples using Agilent SureSelect XT Human All Exon V6 + UTR kit (Agilent Cat. No.5190-8884). Genomic DNA was first cleaved to a size range of approximately 200 to 500 bp, and then sequencing adapters were attached to these fragments. Paired-end sequencing at a read depth of 100X was performed on the exome libraries using the Illumina HiSeq 4000 sequencer.

Analysis of sequences. Bioinformatic analysis of paired-end sequences/reads (FASTQs) started with trimming Illumina adapter sequences in them using Trimmomatic-v0.36 and then verifying the trimmed read lengths and quality using FastQC-v0.11.7[73,74]. All passed the quality check and were aligned to the human reference genome with decoy sequences (human_g1k_v37_decoy.fasta.gz) as provided in the resource package of GATK v.3.7 using Burrow-Wheelers Aligner (BWA-MEM)[75,76]. The output binary alignment map (BAM) files were then processed through Picard-v2.18's MarkDuplicates tool[77]. Base recalibration was performed using the GATK-v3.8's Base Quality Score Recalibration tool[75]. For somatic analysis, the base recalibrated tumor and normal BAMs were fed to Mutect2 (GATK-v3.8)[78] to call somatic variants at chromosome positions covered in the target bed (Agilent SureSelect V6 + UTR). The indel realignment of each tumor and normal BAMs pair were performed using Sentieon Realigner tool-v201711.05[79] on the Seven Bridges Platform before calling variants from VarScan2[80], and Strelka2[81] to reduce the possibility of false SNVs calls arising due to alignment around InDels. GATK's best practices does not recommend performing InDel realignment of BAMs before feeding them to Mutect2 as the tool already performs a better haplotype-based local reassembly of reads as default. However, since Varscan2 lacks such reassembly step and Strelka2 has a rather tier-based local reassembly step that is subject to the evaluation of properties for variant locus, a prior GATK InDel realignment would ensure the improvement in read alignments before they are fed to these two variant callers. The resulting somatic variant call files (VCFs) were hard filtered with read-depth greater than or equal to 50X and variant allele frequency (VAF) greater than or equal to 5% to account for tumor heterogeneity. The read depth threshold of 50X was chosen as that was the minimum average read depth across target bed regions for each tumor sample. At such a considerable read depth threshold in exome sequencing, one can go lower in VAFs to detect somatic variants from tumor samples against matched normal[82]. Somatic variants called by at least two methods were combined and annotated with variant effects using the snpEFF-v4.3t tool[83], and analyzed downstream using OpenCravat-v2.2.2[25,84]. Rare variants that have less than 1% minor allele frequency (MAF) in African population were retained and rest filtered out. The alignment of reads was manually reviewed using BAM files of each tumor and matched normal samples in Integrative Genomics Viewer (IGV) to reduce the risk of false positivity[85,86]. SNV was visualized using Maftools oncoplot function[28] in RStudio version 1.4.1106 by using R 4.1.0. To compare mutation load against TCGA cohorts, tcgaCompare function was used.

Somatic copy number alterations. SCNAs were determined using the CNVKit-v0.9.3 tool in the SevenBridges Genomics interphase[87]. Using CNVkit's Reference module, we created a reference coverage file from the double realigned normal BAMs that were realigned initially among themselves and then, with their respective tumor BAMs. Such a double indel realignment of normal would correct indel alignments at the germline sites that are common between them thereby reducing SNV artifacts around those sites as well and providing a clean data for downstream analysis. With CNVKit Reference dependent workflow, in addition to a Reference coverage, a target bed file of baited regions and an anti-target bed file of regions not included in the target bed were generated. These prerequisite files and re-aligned tumor BAM files were used by the tool to iterate for CNAs in each tumor sample. Further, SCNAs for all tumor samples were aggregated to check for both overlapping and novel SCNAs. CNAPP was used to quantify and plot the frequency of SCNAs and to analyze SCNAs clinical relevance[88]. Focal SCNAs were plotted using Karyoplot in RStudio-v1.1.463[89]. The sequencing readouts of

frequently mutated oncogenes and tumor suppressor genes were manually inspected in the BAM file using the Integrative Genomics Viewer- Broad Institute (IGV)[85,86].

Tumor mutational burden. Tumor mutational burden was calculated using the total number of somatic coding mutations per sample[90].

Microsatellite instability. Microsatellite Instability (MSI) of tumor samples was determined by using MSI sensor in SevenBridges. Cutoff value for MSI is > 3.5%[22]. This analysis is designated for normal-tumor pairs.

Mutational signature profile. In SevenBridges Genomics Interface Mutational Signatures—deconstructSigs 1.8.0 DeconstrucSig[91] was used to identify mutational signatures based on nonnegative matrix factorization by creating a mutational profile and comparing the profile with predefined COSMIC mutation signatures[19]. Default parameters were used in the analysis.

Driver gene prediction. We employed 20/20 + (v1.0.1)[92] and CHASMplus, Cancer Genome Interpreter[93], and dNdScv[94] for driver gene prediction. In 20/20 + analysis, the number of simulations was increased to 100,000 and the 2020plus_100k.Rdata trained classifier was used to improve the prediction performance.

Pathway and protein interaction analysis. Reactome database[95] was used to analyze possible protein interactions among proteins encoded by mutated genes. Overrepresentation results were considered where $p < 0.05$ and FDR < 0.1 thresholds were satisfied.

Regulatory approval and consent for publication. The District of Columbia Veterans Affairs Medical Center Institutional Review Board (DC VAMC IRB #07077) approved this study. Written informed consent and the approval for publication of results was obtained from each patient prior to the procedure.

Data availability

All sequencing data generated or analyzed during this study have been submitted for the National Cancer for Biotechnology Information Sequence Read Archive (SRA) repository with the submission number SUB9998664.

Received: 3 October 2020; Accepted: 24 June 2021
Published online: 20 July 2021

References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2021. *CA Cancer J. Clin.* **71**, 7–33. https://doi.org/10.3322/caac.21654 (2021).
2. Data, S. R. in *SEER Research Data 1975–2016 National Cancer Institute, DCCPS, Surveillance Research Program Based on the November 2018 Submission* (2019).
3. Then, E. O. *et al.* Esophageal cancer: An updated surveillance epidemiology and end results database analysis. *World J. Oncol.* **11**, 55–64. https://doi.org/10.14740/wjon1254 (2020).
4. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7–30. https://doi.org/10.3322/caac.21551 (2020).
5. Miller, J. A., Rege, R. V., Ko, C. Y. & Livingston, E. H. Health care access and poverty do not explain the higher esophageal cancer mortality in African Americans. *Am. J. Surg.* **188**, 22–26. https://doi.org/10.1016/j.amjsurg.2003.12.055 (2004).
6. Wasif, N. *et al.* Racial and socioeconomic differences in the use of high-volume commission on cancer-accredited hospitals for cancer surgery in the United States. *Ann. Surg. Oncol.* **25**, 1116–1125. https://doi.org/10.1245/s10434-018-6374-0 (2018).
7. Abnet, C. C., Arnold, M. & Wei, W. Q. Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* **154**, 360–373. https://doi.org/10.1053/j.gastro.2017.08.023 (2018).
8. Katada, C. *et al.* Alcohol consumption and multiple dysplastic lesions increase risk of squamous cell carcinoma in the esophagus, head, and neck. *Gastroenterology* **151**, 860-869 e867. https://doi.org/10.1053/j.gastro.2016.07.040 (2016).
9. Global Burden of Disease Cancer *et al.* Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study. *JAMA Oncol.* **3**, 524–548. https://doi.org/10.1001/jamaoncol.2016.5688 (2017).
10. Murphy, G. *et al.* International cancer seminars: A focus on esophageal squamous cell carcinoma. *Ann. Oncol.* **28**, 2086–2093. https://doi.org/10.1093/annonc/mdx279 (2017).
11. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175. https://doi.org/10.1038/nature20805 (2017).
12. Agrawal, N. *et al.* Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov.* **2**, 899–905. https://doi.org/10.1158/2159-8290.CD-12-0189 (2012).
13. Bye, H. *et al.* Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis* **33**, 2155–2161. https://doi.org/10.1093/carcin/bgs262 (2012).
14. Deng, J. *et al.* Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. *Nat. Commun.* **8**, 1533. https://doi.org/10.1038/s41467-017-01730-x (2017).
15. Spratt, D. E. *et al.* Racial/ethnic disparities in genomic sequencing. *JAMA Oncol.* **2**, 1070–1074. https://doi.org/10.1001/jamaoncol.2016.1854 (2016).
16. Chen, Z. *et al.* Incidence and survival differences in esophageal cancer among ethnic groups in the United States. *Oncotarget* **8**, 47037–47051. https://doi.org/10.18632/oncotarget.16694 (2017).
17. Pack, S. D. *et al.* Molecular cytogenetic fingerprinting of esophageal squamous cell carcinoma by comparative genomic hybridization reveals a consistent pattern of chromosomal alterations. *Genes Chromosom. Cancer* **25**, 160–168 (1999).
18. Erkizan, H. V. *et al.* African–American esophageal squamous cell carcinoma expression profile reveals dysregulation of stress response and detox networks. *BMC Cancer* **17**, 426. https://doi.org/10.1186/s12885-017-3423-1 (2017).
19. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421. https://doi.org/10.1038/nature12477 (2013).

20. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993. https://doi.org/10.1016/j.cell.2012.04.024 (2012).
21. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407. https://doi.org/10.1038/ng.3441 (2015).
22. Niu, B. *et al.* MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016. https://doi.org/10.1093/bioinformatics/btt755 (2014).
23. Fulda, S. Targeting c-FLICE-like inhibitory protein (CFLAR) in cancer. *Expert Opin. Ther. Targets* **17**, 195–201. https://doi.org/10.1517/14728222.2013.736499 (2013).
24. Karkera, J. D. *et al.* Refinement of regions with allelic loss on chromosome 18p11.2 and 18q12.2 in esophageal squamous cell carcinoma. *Clin. Cancer Res.* **6**, 3565–3569 (2000).
25. Pagel, K. A. *et al.* Integrated informatics analysis of cancer-related variants. *JCO Clin. Cancer Inform.* **4**, 310–317. https://doi.org/10.1200/CCI.19.00132 (2020).
26. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucl. Acids Res.* **47**, D886–D894. https://doi.org/10.1093/nar/gky1016 (2019).
27. Quang, D., Chen, Y. & Xie, X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763. https://doi.org/10.1093/bioinformatics/btu703 (2015).
28. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756. https://doi.org/10.1101/gr.239244.118 (2018).
29. Bavarva, J. H., Tae, H., McIver, L. & Garner, H. R. Nicotine and oxidative stress induced exomic variations are concordant and overrepresented in cancer-associated genes. *Oncotarget* **5**, 4788–4798. https://doi.org/10.18632/oncotarget.2033 (2014).
30. Liu, W. *et al.* Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI Insight* **1**, e88755. https://doi.org/10.1172/jci.insight.88755 (2016).
31. Guo, J. *et al.* Germline and somatic variations influence the somatic mutational signatures of esophageal squamous cell carcinomas in a Chinese population. *BMC Genom.* **19**, 538. https://doi.org/10.1186/s12864-018-4906-4 (2018).
32. Kato, H. *et al.* Gene amplification of EGFR, HER2, FGFR2 and MET in esophageal squamous cell carcinoma. *Int. J. Oncol.* **42**, 1151–1158. https://doi.org/10.3892/ijo.2013.1830 (2013).
33. Duan, X. Z. S., Zhang, M., Wang, P., Zhang, J. & Wang, J. Clinical significance of EGFR and EGFRvIII expression in human esophageal carcinoma. *Pak. J. Med. Sci.* **27**, 490–495 (2011).
34. Pines, G., Huang, P. H., Zwang, Y., White, F. M. & Yarden, Y. EGFRvIV: A previously uncharacterized oncogenic mutant reveals a kinase autoinhibitory mechanism. *Oncogene* **29**, 5850–5860. https://doi.org/10.1038/onc.2010.313 (2010).
35. Gan, H. K., Cvrljevic, A. N. & Johns, T. G. The epidermal growth factor receptor variant III (EGFRvIII): Where wild things are altered. *FEBS J.* **280**, 5350–5370. https://doi.org/10.1111/febs.12393 (2013).
36. Sok, J. C. *et al.* Mutant epidermal growth factor receptor (EGFRvIII) contributes to head and neck cancer growth and resistance to EGFR targeting. *Clin. Cancer Res.* **12**, 5064–5073. https://doi.org/10.1158/1078-0432.CCR-06-0913 (2006).
37. Frederick, L., Wang, X. Y., Eley, G. & James, C. D. Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas. *Cancer Res.* **60**, 1383–1387 (2000).
38. Cui, Y. *et al.* Whole-genome sequencing of 508 patients identifies key molecular features associated with poor prognosis in esophageal squamous cell carcinoma. *Cell Res.* **30**, 902–913. https://doi.org/10.1038/s41422-020-0333-6 (2020).
39. Unni, A. M., Lockwood, W. W., Zejnullahu, K., Lee-Lin, S. Q. & Varmus, H. Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. *Elife* **4**, e06907. https://doi.org/10.7554/eLife.06907 (2015).
40. Khan, S. A., Zeng, Z., Shia, J. & Paty, P. B. EGFR gene amplification and KRAS mutation predict response to combination targeted therapy in metastatic colorectal cancer. *Pathol. Oncol. Res.* **23**, 673–677. https://doi.org/10.1007/s12253-016-0166-2 (2017).
41. Favazza, L. A. *et al.* KRAS amplification in metastatic colon cancer is associated with a history of inflammatory bowel disease and may confer resistance to anti-EGFR therapy. *Mod. Pathol.* **33**, 1832–1843. https://doi.org/10.1038/s41379-020-0560-x (2020).
42. Hewitt, L. C. *et al.* KRAS status is related to histological phenotype in gastric cancer: Results from a large multicentre study. *Gastric Cancer* **22**, 1193–1203. https://doi.org/10.1007/s10120-019-00972-6 (2019).
43. Deng, N. *et al.* A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets. *Gut* **61**, 673–684. https://doi.org/10.1136/gutjnl-2011-301839 (2012).
44. Essakly, A. *et al.* PIK3CA and KRAS amplification in esophageal adenocarcinoma and their impact on the inflammatory tumor microenvironment and prognosis. *Transl. Oncol.* **13**, 157–164. https://doi.org/10.1016/j.tranon.2019.10.013 (2020).
45. Yaeger, R. *et al.* Mechanisms of acquired resistance to BRAF V600E inhibition in colon cancers converge on RAF dimerization and are sensitive to its inhibition. *Cancer Res.* **77**, 6513–6523. https://doi.org/10.1158/0008-5472.CAN-17-0768 (2017).
46. Kim, H. S. *et al.* PIK3CA amplification is associated with poor prognosis among patients with curatively resected esophageal squamous cell carcinoma. *Oncotarget* **7**, 30691–30701. https://doi.org/10.18632/oncotarget.8749 (2016).
47. Kolasa, I. K. *et al.* PIK3CA amplification associates with resistance to chemotherapy in ovarian cancer patients. *Cancer Biol. Ther.* **8**, 21–26. https://doi.org/10.4161/cbt.8.1.7209 (2009).
48. Huang, J. *et al.* Prognostic significance of c-MYC amplification in esophageal squamous cell carcinoma. *Ann. Thorac. Surg.* **107**, 436–443. https://doi.org/10.1016/j.athoracsur.2018.07.077 (2019).
49. Hu, X. *et al.* Amplification and overexpression of CTTN and CCND1 at chromosome 11q13 in esophagus squamous cell carcinoma (ESCC) of North Eastern Chinese Population. *Int. J. Med. Sci.* **13**, 868–874. https://doi.org/10.7150/ijms.16845 (2016).
50. Ying, J. *et al.* Genome-wide screening for genetic alterations in esophageal cancer by aCGH identifies 11q13 amplification oncogenes associated with nodal metastasis. *PLoS ONE* **7**, e39797. https://doi.org/10.1371/journal.pone.0039797 (2012).
51. Campbell, J. D. *et al.* Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell. Rep.* **23**, 194-212 e196. https://doi.org/10.1016/j.celrep.2018.03.063 (2018).
52. Luen, S. J. *et al.* Association of somatic driver alterations with prognosis in postmenopausal, hormone receptor-positive, HER2-negative early breast cancer: A secondary analysis of the BIG 1–98 randomized clinical trial. *JAMA Oncol.* **4**, 1335–1343. https://doi.org/10.1001/jamaoncol.2018.1778 (2018).
53. Hermida-Prado, F. *et al.* Distinctive expression and amplification of genes at 11q13 in relation to HPV status with impact on survival in head and neck cancer patients. *J. Clin. Med.* https://doi.org/10.3390/jcm7120501 (2018).
54. Ramos-Garcia, P. *et al.* An update of knowledge on cortactin as a metastatic driver and potential therapeutic target in oral squamous cell carcinoma. *Oral Dis.* **25**, 949–971. https://doi.org/10.1111/odi.12913 (2019).
55. Zucman-Rossi, J., Villanueva, A., Nault, J. C. & Llovet, J. M. Genetic landscape and biomarkers of hepatocellular carcinoma. *Gastroenterology* **149**, 1226-1239 e1224. https://doi.org/10.1053/j.gastro.2015.05.061 (2015).
56. Gollin, S. M. Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: A next generation window to the biology of disease. *Genes Chromosom. Cancer* **53**, 972–990. https://doi.org/10.1002/gcc.22214 (2014).
57. Brown, J., Stepien, A. J. & Willem, P. Landscape of copy number aberrations in esophageal squamous cell carcinoma from a high endemic region of South Africa. *BMC Cancer* **20**, 281. https://doi.org/10.1186/s12885-020-06788-3 (2020).
58. Hu, N. *et al.* Genome-wide loss of heterozygosity and copy number alteration in esophageal squamous cell carcinoma using the Affymetrix GeneChip Mapping 10 K array. *BMC Genom.* **7**, 299. https://doi.org/10.1186/1471-2164-7-299 (2006).

59. Yen, C. C. *et al.* Comparative genomic hybridization of esophageal squamous cell carcinoma: Correlations between chromosomal aberrations and disease progression/prognosis. *Cancer* **92**, 2769–2777. https://doi.org/10.1002/1097-0142(20011201)92:11%3c2769::aid-cncr10118%3e3.0.co;2-m (2001).
60. Sondka, Z. *et al.* The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705. https://doi.org/10.1038/s41568-018-0060-1 (2018).
61. Ma, S. *et al.* Targeted therapy of esophageal squamous cell carcinoma: The NRF2 signaling pathway as target. *Ann. N. Y. Acad. Sci.* **1434**, 164–172. https://doi.org/10.1111/nyas.13681 (2018).
62. Chang, J. *et al.* Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat. Commun.* **8**, 15290. https://doi.org/10.1038/ncomms15290 (2017).
63. Wang, K. *et al.* Comprehensive genomic profiling of advanced esophageal squamous cell carcinomas and esophageal adenocarcinomas reveals similarities and differences. *Oncologist* **20**, 1132–1139. https://doi.org/10.1634/theoncologist.2015-0156 (2015).
64. Letouze, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315. https://doi.org/10.1038/s41467-017-01358-x (2017).
65. Hatakeyama, K. *et al.* Mutational burden and signatures in 4000 Japanese cancers provide insights into tumorigenesis and response to therapy. *Cancer Sci.* **110**, 2620–2628. https://doi.org/10.1111/cas.14087 (2019).
66. Lin, D. C. *et al.* Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut* **67**, 1769–1779. https://doi.org/10.1136/gutjnl-2017-314607 (2018).
67. Li, X. C. *et al.* A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma. *Ann. Oncol.* **29**, 938–944. https://doi.org/10.1093/annonc/mdy011 (2018).
68. Wei, R. *et al.* Comprehensive analysis reveals distinct mutational signature and its mechanistic insights of alcohol consumption in human cancers. *Brief Bioinform.* https://doi.org/10.1093/bib/bbaa066 (2021).
69. Plath, M. *et al.* Unraveling most abundant mutational signatures in head and neck cancer. *Int. J. Cancer* **148**, 115–127. https://doi.org/10.1002/ijc.33297 (2021).
70. Zhang, L. *et al.* Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.* **96**, 597–611. https://doi.org/10.1016/j.ajhg.2015.02.017 (2015).
71. Hao, J. J. *et al.* Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat. Genet.* **48**, 1500–1507. https://doi.org/10.1038/ng.3683 (2016).
72. Edge, S. B. & Compton, C. C. The American joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* **17**, 1471–1474. https://doi.org/10.1245/s10434-010-0985-4 (2010).
73. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
75. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498. https://doi.org/10.1038/ng.806 (2011).
76. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11–33. https://doi.org/10.1002/0471250953.bi1110s43 (2013).
77. Broad Institute. *Picard* http://broadinstitute.github.io/picard/index.html.
78. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219. https://doi.org/10.1038/nbt.2514 (2013).
79. Freed, D. N., Weber, J. A. & Edwards, J. S. The sentieon genomics tools—A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv* https://doi.org/10.1101/115717 (2017).
80. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576. https://doi.org/10.1101/gr.129684.111 (2012).
81. Kim, S. *et al.* Strelka2: Fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594. https://doi.org/10.1038/s41592-018-0051-x (2018).
82. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91. https://doi.org/10.1186/s13073-020-00791-w (2020).
83. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92. https://doi.org/10.4161/fly.19695 (2012).
84. Kim, S. Y., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. *BMC Bioinform.* **15**, 154. https://doi.org/10.1186/1471-2105-15-154 (2014).
85. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. https://doi.org/10.1038/nbt.1754 (2011).
86. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192. https://doi.org/10.1093/bib/bbs017 (2013).
87. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873. https://doi.org/10.1371/journal.pcbi.1004873 (2016).
88. Franch-Exposito, S. *et al.* CNApp, a tool for the quantification of copy number alterations and integrative analysis revealing clinical implications. *Elife* https://doi.org/10.7554/eLife.50267 (2020).
89. Gel, B. & Serra, E. karyoploteR: An R/bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090. https://doi.org/10.1093/bioinformatics/btx346 (2017).
90. Zang, Y. S. *et al.* Comprehensive analysis of potential immunotherapy genomic biomarkers in 1000 Chinese patients with cancer. *Cancer Med.* **8**, 4699–4708. https://doi.org/10.1002/cam4.2381 (2019).
91. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31. https://doi.org/10.1186/s13059-016-0893-4 (2016).
92. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14330–14335. https://doi.org/10.1073/pnas.1616440113 (2016).
93. Tamborero, D. *et al.* Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25. https://doi.org/10.1186/s13073-018-0531-8 (2018).
94. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029-1041 e1021. https://doi.org/10.1016/j.cell.2017.09.042 (2017).
95. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucl. Acids Res.* **48**, D498–D503. https://doi.org/10.1093/nar/gkz1031 (2020).

Acknowledgements

This Project has been funded (to HVE) in whole or in part with Federal funds (UL1TR000101 previously UL1RR031975) from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through the Clinical and Translational Science Awards Program (CTSA), a trademark of DHHS,

part of the Roadmap Initiative, "Re-Engineering the Clinical Research Enterprise." We would like to thank the Institute for Clinical Research, Washington DC VA Medical Center for their support.

Author contributions

Conceptualization: H.V.E., V.N., R.W., Design of bioinformatic analysis, bioinformatic analyses, and interpretation of data: S.S., H.V.E., T.G.N., Pathological analysis: J.H.L., Sampling biopsy materials: G.M., Manuscript preparation: H.V.E. drafted and all authors contributed to the writing of the manuscript. Final approval of manuscript: All authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-94064-0.

Correspondence and requests for materials should be addressed to H.V.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.