nature portfolio

https://doi.org/10.1038/s41588-021-00928-6

Supplementary information

Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence

In the format provided by the authors and unedited

Supplementary Information

Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence

Author List

Sarah Moody[1], Sergey Senkin[2], S M Ashiqul Islam[3,4,5], Jingwei Wang[1], Dariush Nasrollahzadeh[6,2], Ricardo Cortez Cardoso Penha[2], Stephen Fitzgerald[1], Erik N Bergstrom[3,4,5], Joshua Atkins[2], Yudou He[3,4,5], Azhar Khandekar[3,4,5], Karl Smith-Byrne[2], Christine Carreira[7], Valerie Gaborieau[2], Calli Latimer[1], Emily Thomas[1], Irina Abnizova[1], Pauline E Bucciarelli[1], David Jones[1], Jon W Teague[1], Behnoush Abedi-Ardekani[2], Stefano Serra[8], Jean-Yves Scoazec[9], Hiva Saffar[10], Farid Azmoudeh-Ardalan[11], Masoud Sotoudeh[6], Arash Nikmanesh[6], Hossein Poustchi[6], Ahmadreza Niavarani[6], Samad Gharavi[6], Michael Eden[12], Paul Richman[13], Lia S Campos[14], Rebecca C Fitzgerald[15], Luis Felipe Ribeiro[16], Sheila Coelho Soares-Lima[16], Charles Dzamalala[17], Blandina Theophil Mmbaga[18], Tatsuhiro Shibata[19], Diana Menya[20], Alisa M Goldstein[21], Nan Hu[21], Reza Malekzadeh[6], Abdolreza Fazel[22], Valerie McCormack[23], James McKay[2], Sandra Perdomo[2], Ghislaine Scelo[2,24], Estelle Chanudet[2], Laura Humphreys[1], Ludmil B Alexandrov[3,4,5], Paul Brennan[2], Michael R Stratton[1]

These authors contributed equally: Sarah Moody, Sergey Senkin

* Corresponding author: Michael R Stratton (mrs@sanger.ac.uk)

Affiliations

[1]Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK, [2]Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), 150 Cours Albert Thomas, Lyon, 69372, France, [3]Moores Cancer Centre, UC San Diego Health, La Jolla, California, 92037, USA, [4]Department of Cellular and

Molecular Medicine, University of California, La Jolla, California, 92093, USA, [5]Department of Bioengineering, University of California, La Jolla, California, 92093, USA, [6]Digestive Oncology Research Center, Digestive Diseases Research Institute, Tehran University of Medical Sciences, Shariati Hospital, Tehran, Iran, [7]Evidence Synthesis and Classification Branch, International Agency for Research on Cancer (IARC/WHO), 150 Cours Albert Thomas, Lyon, 69372, France, [8]University Health Network, Toronto, Canada, [9]Department Laboratory Medicine and Pathology, Gustave Roussy, Paris, France, [10]Department of Pathology, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran, [11]Liver Transplantation Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Tehran, Iran, [12]Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK, [13]Histopathology Department, Hemel Hempstead General Hospital, Hemel Hempstead, HP2 4AD, UK, [14]West Suffolk NHS Foundation Trust, Bury St Edmunds, IP33 2QZ, UK, [15]MRC Cancer Unit, University of Cambridge, Cambridge, CB2 0XZ, UK, [16]Brazilian National Cancer Institute, Rio de Janeiro, Brazil, [17]University of Malawi, Malawi, [18]Kilimanjaro Clinical Research Institute, Kilimanjaro Christian Medical Centre & Kilimanjaro Christian Medical University College, Tanzania, [19]Division of Cancer Genomics, National Cancer Centre Research Institute, Tokyo, Japan, [20]Moi University, Kenya, [21]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, USA, [22]Golestan Research Center of Gastroenterology and Hepatology, Golestan University of Medical Sciences, Gorgan, Iran, [23]Environment and Lifestyle Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), 150 Cours Albert Thomas, Lyon, 69372, France, [24]Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Italy

Supplementary Information Inventory

This file contains the following:

Supplementary Note

A) Supplementary Results

Copy Number Profiles

Whilst this dataset was not ideal to investigate copy number changes as the cohort is comprised of cases with variable purity, which impacts the reliability of copy number profile calling, the copy number profiles were investigated in a subset of cases with tumour purity >50%. As expected, overall the copy number profiles in ESCC were complex, with no significant differences in the number of CNV segments detected (Supplementary Fig. 4a/b, Supplementary Table 1).

Novel *de novo* SBS and DBS signatures

One SBS288 *de novo* signature (SBS288P) could not be decomposed into COSMIC reference signatures, and was found in 30/552 (5%) cases. This signature was present in the HDP results (Extended Data Fig. 2a) and has also been identified in stomach cancers from a recent reanalysis of the pan cancer cohort used to derive the COSMIC reference signatures, but the etiology remains unknown[1]. SBS288P was most frequent in China (15/138, 11%), followed by Malawi (3/59, 5%), Kenya (3/68, 4%), Iran (8/178, 4%) and Brazil (1/30, 1%), but was not observed in Japan, Tanzania or the UK, most likely due to the small size of these cohorts.

The novel DBS *de novo* signature DBS78D was present in 64% of cases with an average contribution of 26% to the mutation burden. This signature could not be decomposed, but contained CC>NN peaks consistent with COSMIC reference DBS11, which is thought to be associated with APOBEC mutagenesis, in addition to TC>NN peaks which whilst present in DBS11, were more prominent in the novel extracted signature[2].

Structural rearrangement signatures

The number of rearrangements ranged from 1 – 1110 (median 218) per case, with no differences between countries observed (Supplementary Fig. 5a). Structural rearrangement signatures were extracted using the 32 subclasses previously described[3]. This resulted in ten *de novo* signatures (Supplementary Fig. 5b-d, Supplementary Table 5,8), which could not be further decomposed as reference rearrangement signatures do not currently exist. The dominant signature was RS32A, which was defined by non-clustered translocations. This signature was present in 98% of cases and on average accounted for 33% of the rearrangement burden. The remaining signatures accounted for between 4-11% of the rearrangement burden on average. RS32A, RS32B and RS32I were similar to breast cancer rearrangement signatures RS2, RS4 and RS1 respectively (Supplementary Table 7), whereas the other signatures did not closely resemble those previously reported[3]. No significant differences in the attributions of any of the rearrangement signatures was observed in any individual country. (Supplementary Fig. 5c, Supplementary Tables 8, 10).

Rearrangement signatures were included in the regression analysis; however, no signatures were found between and ESCC risk factors and any of the *de novo* signatures. Despite previous studies finding associations between cases with BRCA1/2 variants and structural rearrangement signatures in breast cancer, no significant associations were found in ESCC[3].

Clustered mutations

The number and type of clustered mutations were investigated in the subset of cases with tumor purity >50% (Methods). This included doublet substitutions, the recently defined omikli events[4] and kataegis events, which collectively represented an average 3.24% (0.82%, 1.96% and 0.46% respectively) of the total SBS mutation burden. There was no significant difference between countries in either the total number or proportion of any clustered mutation type (Supplementary Fig.6).

B) Supplementary Methods

Comparison of Mutational Patterns from Different Variant Calling Approaches

In this study we used the combination of CaVEMan + Strelka2 and Pindel + Strelka2 for the final set of single base substitutions and indels. In order to examine whether the choice of second caller had a significant impact on the mutational profiles, 10 evaluated samples were compared with different variant call approaches (Supplementary Methods Fig.7), including GATK4 Mutect2[5], Strelka2[6], Varscan2[7] and MuSE[8]. Note that MuSE only identifies single base substitutions while the other three tools also provide small insertions and deletions. For Mutect2, paired reads were allowed to independently support different haplotypes during initial variant calling and the expected frequency of alleles not found in the germline resource was 0.00003125 (considering 16,000 as the number of protein coding genes known in the human genome). Contamination table and read orientation models were built from the paired samples and were subsequently used for filtering. For VarScan2, the initial variant calling expected a tumor purity of 0.8 and the subsequent filtering required a minimum coverage of 10 reads and at least 3 alternative reads in tumor with a minimum alternative allele frequency of 0.20. For Strelka2 and MuSE, the default setting for whole genome sequencing were used to produce a list of raw and filtered variants. Cosine similarities between the SBS96 contexts of single base substitutions were calculated by comparing the reference to mutations from (i) Mutect2, (ii) Strelka2, (iii) Varscan2, (iv) MuSE, (v) CaVEMan +

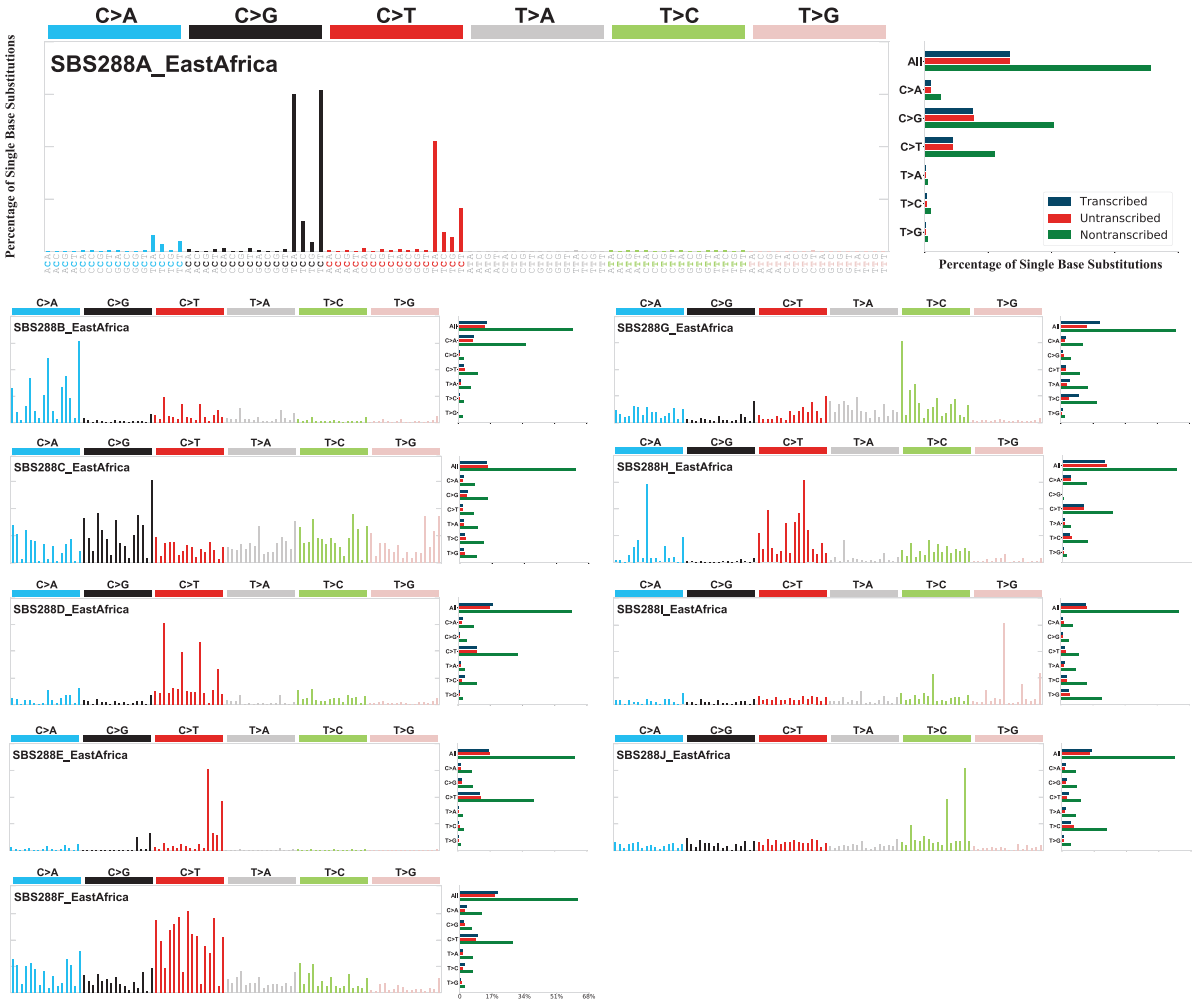
Mutect2, (vi) CaVEMan + Varscan2, (vii) CaVEMan + MuSE. Similarly, cosine similarities between the ID83 contexts of small insertions and deletions were calculated by comparing the reference to mutations from (i) Mutect2, (ii) Strelka2, (iii) Varscan2, (iv) CaVEMan + Mutect2, (v) CaVEMan + Varscan2. The results showed that regardless of the second caller used, the cosine similarities of the mutational spectra generated were very similar (Supplementary Fig4 a/b).

Supplementary Figures

Supplementary Fig.1: Regional SBS288 signature extractions - China

Supplementary Fig.1: Regional SBS288 signature extractions – China. Ten *de novo* signatures extracted from 138 ESCC cases from Shanxi province, China

Supplementary Fig.2: Regional SBS288 signature extractions – East Africa

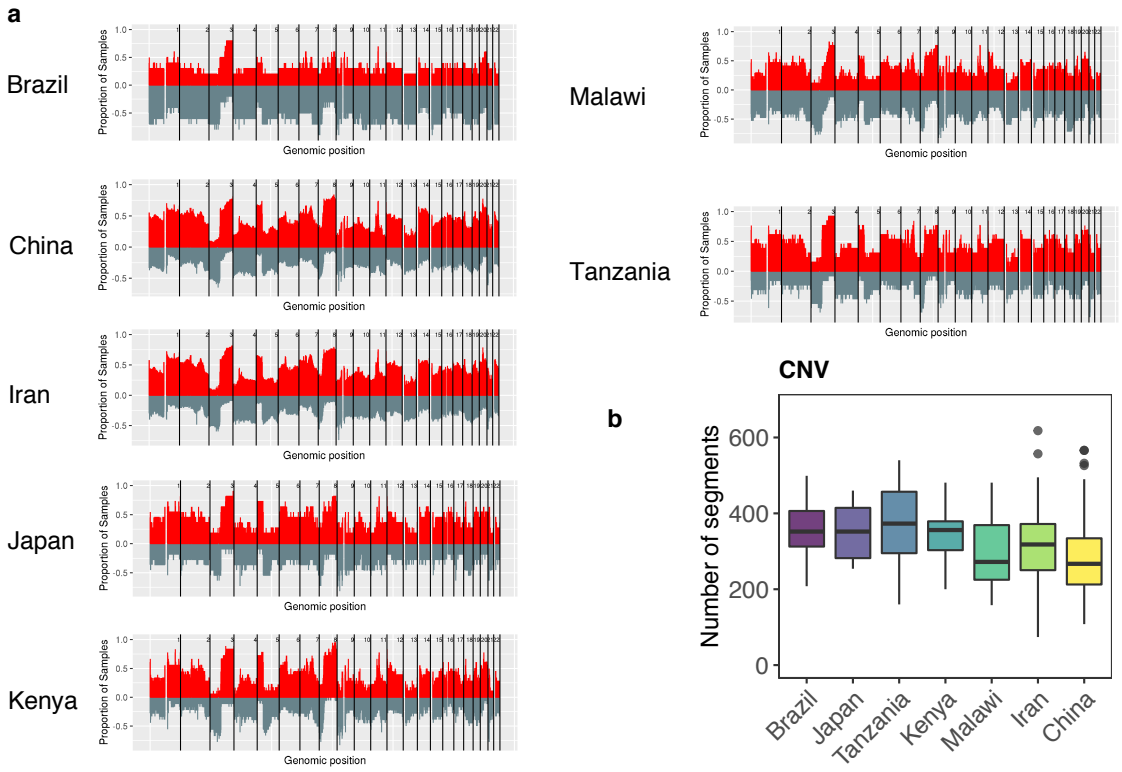


Supplementary Fig.2: Regional SBS288 signature extractions – East Africa. Ten de novo signatures extracted from 162 ESCC cases from Kenya, Tanzania and Malawi.

Supplementary Fig.3: Regional SBS288 signature extractions - Iran

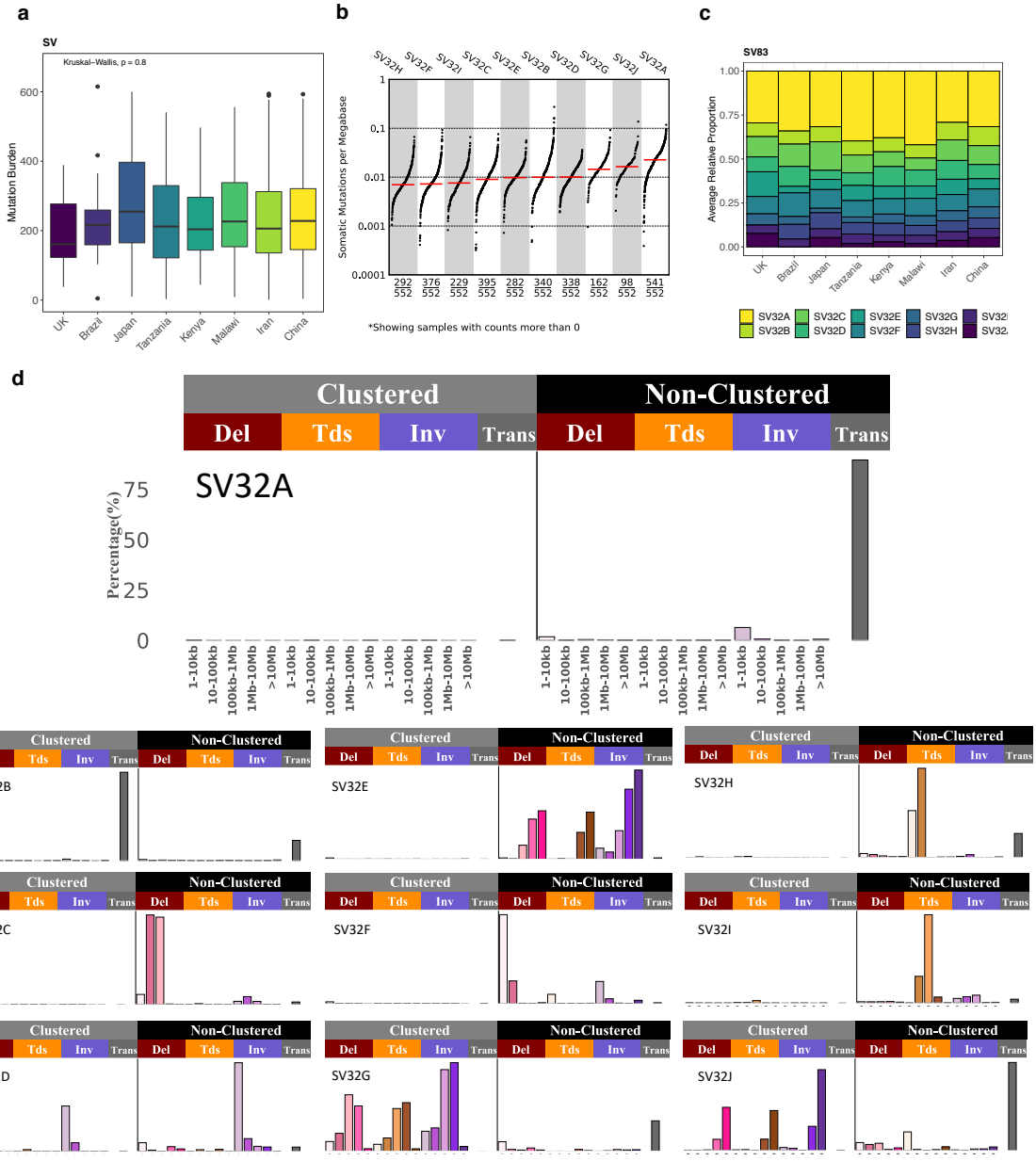
Supplementary Fig.3: Regional SBS288 signature extractions – Iran. Thirteen de novo signatures extracted from 178 ESCC cases from Golestan province, Iran.

Supplementary Fig.4: Copy number profiles in ESCC



Supplementary Fig.4: Copy number profiles in ESCC. (a) Complex patterns of gains and loss in 185 ESCC tumor genomes. Cases with tumor purity <50% were excluded from this analysis, and UK cases were not included as only two cases were of sufficient purity. (b) Number of CNV segments per case in as determined by the ASCAT algorithm, showing no significant difference between countries (Kruskal-Wallis (two-sided) test, n=185 biologically independent samples). Cases with tumor purity <50% were excluded from this analysis, and UK cases were not included as only two cases were of sufficient purity. Box and whiskers plots are in the style of Tukey. Cases with tumor purity <50% were excluded from this analysis, and UK cases were not included as only two cases were of sufficient purity. The line within the box is plotted at the median while the upper and lower ends are indicated 25th and 75th percentiles. Whiskers show 1.5*IQR (interquartile range) and values outside it are shown as individual data points. Countries are ordered by approximate incidence rate in ascending order.

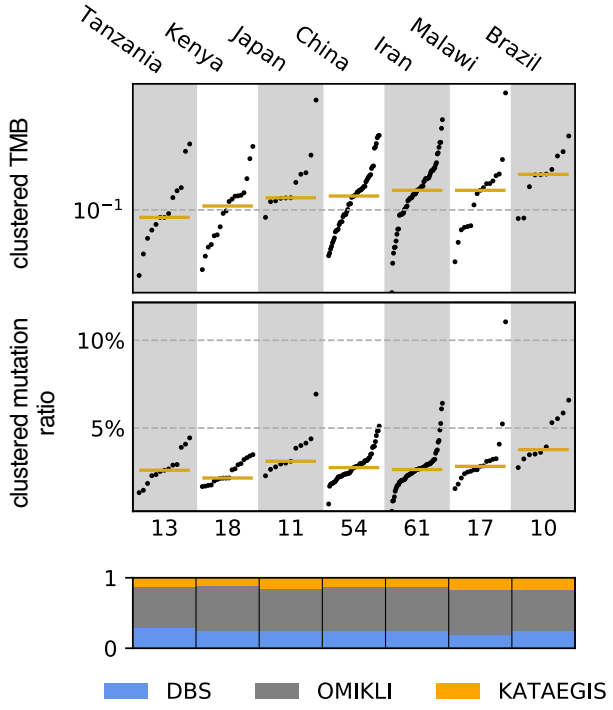
Supplementary Fig.5: SV mutational signature analysis of ESCC



Supplementary Fig.5: SV32 mutational signature analysis of ESCC. (a) No significant difference was observed in the number of rearrangements between countries (Kruskal-Wallis (two-sided) test, n=552 biologically independent samples). Hypermutators were defined using the interquartile range method, with cases with mutation burdens more than 1.5 IQR above Q3 removed from the analysis. Box and whiskers plots are in the style of Tukey. The line within the

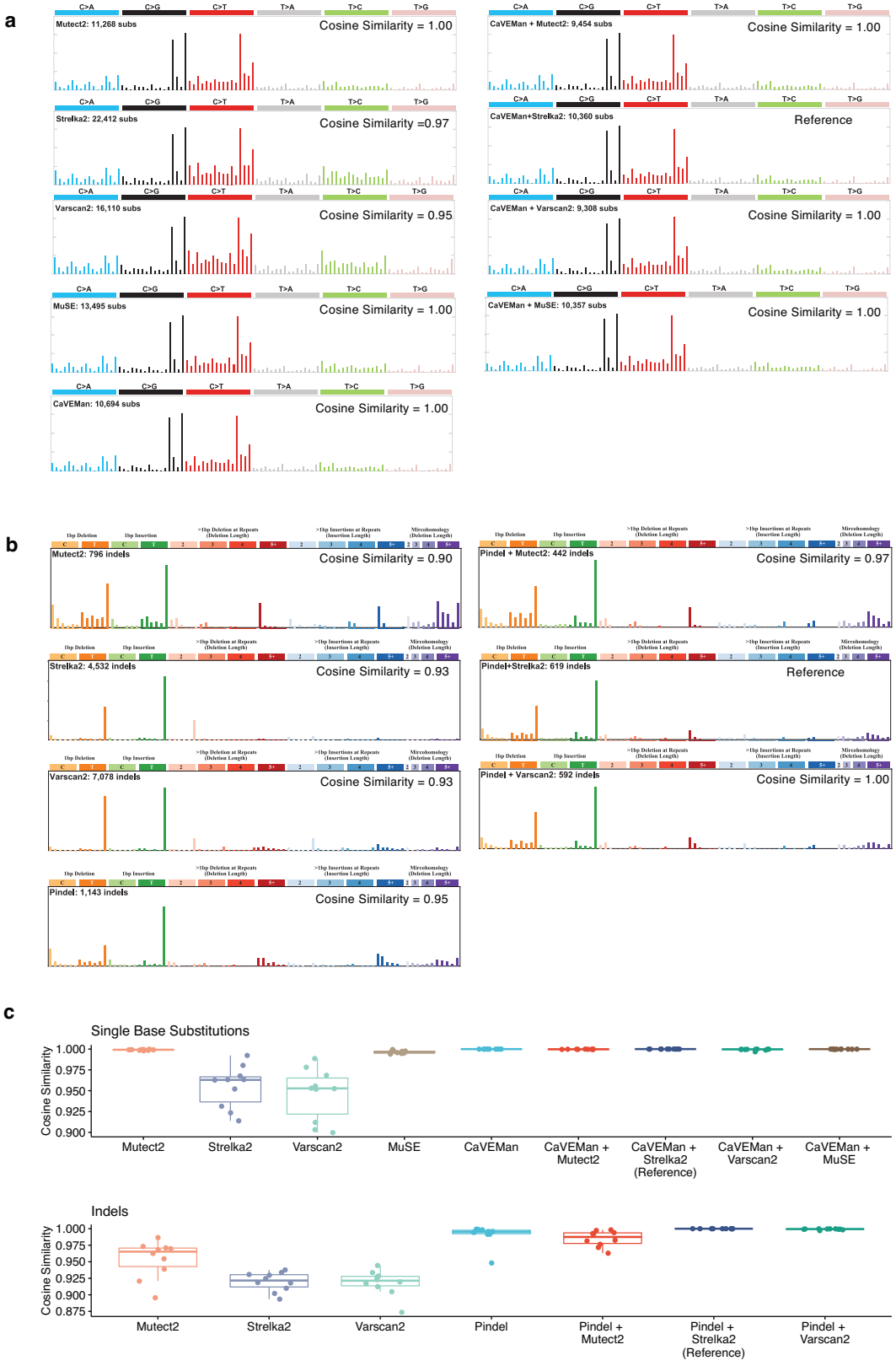
box is plotted at the median while the upper and lower ends are indicated 25th and 75th percentiles. Whiskers show 1.5*IQR (interquartile range) and values outside it are shown as individual data points. Countries are ordered by approximate incidence rate in ascending order. (b) TMB plot showing the frequency and mutations/mb for each of the extracted SV32 *de novo* signatures. (c) Average relative attributions of SV *de novo* signatures are broadly similar between all countries. Countries are ordered by approximate incidence rate in ascending order. (d) Ten *de novo* signatures extracted from 552 ESCC cases.

Supplementary Fig.6: Clustered mutations in ESCC



Supplementary Fig.6: Clustered mutations in ESCC. The distribution of clustered mutations in ESCC compared across each country ordered by median tumor mutational burden (TMB), where each dot represents a single tumor. The clustered mutation ratio is calculated as the fraction of clustered mutations compared to the total number of mutations in a given sample. Each clustered event is subclassified and summarized as the proportion of mutations per country associated with a double-base substitution event, an omikli event, or as a kataegis event.

Supplementary Fig 7. Comparison of Mutational Patterns from Different Variant Calling Approaches



Supplementary Fig 7. Comparison of Mutational Patterns from Different Variant Calling Approaches. An example comparing the patterns of somatic mutations for a single ESCC case is shown for both single base substitutions (a) and small insertions and deletions (b). The mutational profiles of single base substitutions are highly stable across different variant callers in this sample (a) as well as across all 10 examined samples (c). Specifically, when compared to the reference, Mutect2, MuSE and all 2-caller approaches produced cosine similarity > 0.99 while Strelka2 and Varscan2 each showed patterns with median cosine similarity > 0.95 (c). The mutational profile of small insertions and deletions were less stable in this case and across all other examined cases. When compared to the reference, Mutect2, Strelka2 and Varscan2 generated mutational patterns with median cosine similarities between 0.92 and 0.97. On the contrary, MuSE and all of the 2-callers yielded highly similar spectrums (cosine similarity close to 1).

Supplementary References

1. Ashiqul Islam, S. M. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *bioRxiv* 2020.12.13.422570 (2020).
2. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
3. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
4. Mas-Ponte, D. & Supek, F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat. Genet.* **52**, 958–968 (2020).
5. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *BioRxiv* (2019).
6. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594 (2018).

7. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-76 (2012).
8. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* **17**, 178 (2016).