nature research

Corresponding author(s): Mike Stratton

Last updated by author(s): Jun 15, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Whole genome sequencing (150bp paired end) was performed on the Illumina HiSeqX platform with target coverage of 40X for tumors and 20X for paired blood. REDCap 10.0.23 was used to collect epidemiological data for cases from Cambridge and Gorgan
Data analysis	Algorithms used: Variant Calling: (https://github.com/cancerit) BWA-Mem v0.7.16a and v0.7.17 ASCAT v4.0.1, v4.1.2 and v4.3.3 BATTENBERG v3.2.2, v3.3.0 and v3.3.1 cgpCaVEMan v1.13.14 and v1.14.1 cgpPINDEL v3.1.2, v3.2.0, v3.2.2 and v3.3.0 BRASS v6.2.0 and v6.3.0 Strelka2 v2.9.10 and Manta 1.6.0 Conpair v0.2 (https://github.com/nygenome/Conpair) SigProfilerMatrixGenerator v1.1.15 (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator) SigProfilerExtractor v1.0.14: (https://github.com/AlexandrovLab/SigProfilerExtractor) MSA v1.1 (https://gitlab.com/s.senkin/MSA) SnpEff 4.3t (https://pcingola.github.io/SnpEff/) HDP v0.1.5 (https://github.com/nicolaroberts/hdp) PLINK v2.00a (www.cog-genomics.org/plink/2.0/) CHORD v2.0 (https://github.com/UMCUGenetics/CHORD)

Statistical analysis was performed in R version 3.6.3 and Python version 3.7.9

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Whole genome sequencing data and patient metadata are deposited in the European Genome 715 Phenome Archive (EGA) associated with study EGAS00001002725. BAM files for all cases included in the final analysis were deposited in dataset EGAD00001006868, and patient metadata in dataset EGAD00001006732. The human reference genome used for alignment is available at ftp://ftp.sanger.ac.uk/pub/cancer/support-files/reference/GRCh37d5.fa

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Cases were selected from retrospective studies from populations which reflect a range of ESCC incidence rates. Numbers were limited by the number of cases available
Data exclusions	Cases were excluded for any of the following pre-established criteria; 1) Incomplete data on core set variables (age at diagnosis, sex, alcohol and tobacco consumption) 2) Failure to pass pathology review as described in the Methods 3) If matched tumour/normal tissue did not originate from the same individual as determined by Fluidigm SNP genotyping. 4) If sequencing coverage was below 30X for tumour, or 15X for matched normal tissue 5) if contamination level was above 3% as determined by Conpair
Replication	Signature analysis using SigProfilerExtractor was replicated independently at both Wellcome Sanger Institute and UCSD (Signature Extraction), and IARC (Signature Attribution) to ensure consistency. Regression analysis was replicated independently with statistical packages in R, Stata and python. All attempts at replication were successful. No other experiments other than those mentioned here were replicated independently.
Randomization	Randomization is not relevant for this study. Cases did not undergo interventions. All cases were collected based on diagnosis of primary squamous cell carcinoma of the esophagus (ESCC) and no prior treatment.
Blinding	Blinding is not relevant for this study. Cases were not subject to any interventions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	552 cases (216 women and 336 men) diagnosed with ESCC were included from the following countries: Brazil (30), China (138), Iran (178), Japan (37), Kenya (68), Malawi (59), Tanzania (35) and United Kingdom (7). Age of diagnosis ranging from 23 to 93 y.o., mean (SD): 60.9 (11.5). All patient information is available in the Patient Metadata available in EGA dataset EGAD00001006732
Recruitment	IARC/WHO coordinated cases recruitment through an international network of collaborators in 8 countries. The inclusion criteria for patients were >=18 years of age, confirmed diagnosis of primary squamous cell carcinoma of the esophagus (ESCC) and no prior treatment. Patients were excluded if they had any condition that could interfere with their ability to provide informed consent or if there were no means of obtaining adequate tissue/ blood samples as per the protocol requirements. The authors are not aware of any potential self-selection bias or other biases present
Ethics oversight	Ethical approvals were obtained from each Local Research Ethics Committee and Federal Ethics Committee as listed below. The study was submitted and approved by the IARC Ethics Committee. Informed consent was obtained from all participants Tanzania: National Institute for Medical Research Tanzania Kenya: Institutional Research and Ethics Committee of Moi University Iran: Tehran University of Medical Science - Vice-Chancellor in Research Affairs UK: Cambridgeshire 4 Research Ethics Committee Brazil: Comitê de Ética em Pesquisa - Instituto Nacional de Câncer Japan: National Research Council - National Cancer Research Center - Research Ethics Review Committee China: Shanxi Cancer Hospital and Institute Malawi: Queen Elizabeth Central Hospital in Malawi -Malawi National Health Sciences Research Committee (NHSRC)

Note that full information on the approval of the study protocol must also be provided in the manuscript.