nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence
Corresponding author name(s): Professor Michael Stratton

Reviewer Comments & Decisions:

Decision Letter, initial version:

18th Jan 2021

Dear Professor Stratton,

First of all, happy new year. I hope you had a restful break - it feels like quite a long time ago now.

Your Article, "Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

Reviewer #1 (oesophageal cancer genomics) considers the paper to be technically robust. Their main concern is with the overall conceptual advance. Please be assured that we are comfortable with the novelty afforded by the work and are therefore prepared to overrule on this point. They raise some queries regarding the analysis, they ask for the provision of missing experimental details and expanded narrative of some key talking points. These should be addressed in full.

Reviewer #2 (cancer genomics inc. trans-ancestry) is broadly supportive of the work but asks for some additional analysis to strengthen and clarify. Similar to Reviewer #1, they have asked you to fill in some missing details regarding methodology. Please address all their comments in full.

Reviewer #3 (genetic epidemiology) also asks for more details regarding methodology, including more information pertaining to your cohorts. More broadly, they ask for expanded discussion to place your findings in the context of previous studies. These comments should be addressed in full. You'll also see that they've asked for further validation regarding the novel signature identified in the Iranian cohort. Having considered this request, we think that you are very welcome to include these data if you have them. But we their absence would not significantly undermine the key message of the paper (in our opinion), and would not thus preclude our editorial interest in the work. However, please do address the comment textually in your rebuttal letter.

Your paper is effectively reporting a negative result (albeit a most intriguing one), which is reminiscent to the study by Riva et al. which we recently published (https://doi.org/10.1038/s41588-020-0692-4). That work, as you know, was performed in mice, but together these papers raise interesting questions about the broader use of mutational signatures to identify environmental exposures. We ask that you explore this further in your Discussion. For example, we'd be interested in an expanded narrative of the sensitivity of this type of analysis; what are your limits of detection? How many mutations are required to detect a signature? How different would exposures need to be to see a quantifiable difference in mutational signatures? These were questions raised during our own editorial discussions about the paper, and we think that a broader commentary of these points would be valuable.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Safia Danovi
Editor
Nature Genetics

Referee expertise:

Referee #1: Oesophageal cancer genomics inc. mutational signatures

Referee #2: Cancer genomics inc. trans-ancestry analyses

Referee #3: Cancer genetic epidemiology inc. gene/environmental interactions

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The authors present an overview of mutational and structural variant signatures inferred from 552

ESCC genomes which have been collected from multiple high and low incidence regions, and relate these patterns with epidemiological parameters in view of identifying potential new exogenous sources for the observed mutational signatures.

The study benefits from well curated data acquired through an epidemiological questionnaire (not extremely "extensive", but quantifying useful parameters). The authors have been very thorough regarding their approaches for inferring the mutational signatures: they test and compare multiple variant callers, and have also tried two different mutational signature methods, even though the latter is not discussed in great detail. This gives confidence that the mutational signatures inferred are as robust as one might expect them to be given the usual caveats. The statistics employed are appropriate for this type of study.

The study recapitulates epidemiological associations with mutational signatures in ESCC which had been reported in other studies, it highlights a role for DDR deficiencies in ESCC (with BRCA1/2 variants observed in a subset of cases) and it identifies a potential new signature which may be linked to opium consumption (however, not clearly distinguishable as stand-alone). Appropriate credit was given to previous work. Most of the results are summarised clearly, but the figures are not always the best suited to illustrate differences between groups (as specified in my comments below).

This is a study with robust methodology, but overall seems to lack the novelty that would be expected for a Nature Genetics publication. The claimed aspects of novelty are preliminary and rather speculative in nature.

Major comments and suggested improvements:
1. HDP was run in addition to SigProfiler to infer mutational signatures. While HDP results are briefly presented in the extended figures, they are difficult to compare to the ones from SigProfiler. It would be useful if the authors could generate some summary statistics or visualisations that highlight to what extent the results from the two methods agree.
2. The authors briefly mention a signature SBS288P of unknown aetiology which has also been reported in stomach cancers and is present in 5% of the samples. Could the authors clarify which countries this signature has been observed in and at what prevalence in each? Was this signature also captured using HDP?
3. The authors investigated copy number profiles "in a subset of cases" (line 124). It is not clear why copy number variation could not be investigated in the entire cohort. The average copy number profiles were presented by country and indeed they look similar, but it is difficult to quantify from this visualisation whether any regions may differ significantly. Beyond the total number of segments altered, are there specific regions that are systematically changed (e.g. amplified or deleted) by country? Were there any cancer drivers specifically amplified or deleted only in certain countries?
4. The authors analyse the TP53 mutation spectrum and compare it between smokers/non-smokers and drinkers/non-drinkers. It was not clear to me whether the authors are looking here at mutations acquired only within the TP53 gene itself or at something else (TP53 mutated vs wild type cancers)? It would be useful if the authors clarified how they derive the spectrum in Figure 4a. Also, an enrichment of SBS16 mutations is reported in drinkers, but the plots in Figure 4 are not helpful in highlighting this. In fact, I do not think any of the panels in Figure 4 are particularly helpful in illustrating the claims the authors are making in the paragraph spanning lines 238-249. Could the authors find an alternative, more quantifiable visualisation in addition to the mutational spectra provided?
5. The authors stress on the importance of the APOBEC mutational process in the development of ESCC. While it is clear this is likely to have a strong contribution, this has been reported before and it

does not seem the clonality analysis in bulk samples of varying purity would be of high enough confidence to make any claims about the timing of APOBEC mutagenesis (even though it is useful to include this as a minor point). The non-significant results in Figure 5 would have fit better in the supplementary material, and do not justify it being a standalone figure by itself.
6. The authors speculate APOBEC mutagenesis might be a mandatory step for ESCC oncogenesis – however, there is not enough evidence to support that it acts with certainty in all cases, or its link to TP53 loss of function. From the figures presented, it was difficult to estimate the exact number of cases with an APOBEC signature, and the overall prevalence of this signature across all the cases in all cohorts (beyond country-level averages). How do the authors explain that other studies have only found evidence for APOBEC mutagenesis only in a subset of tumours (e.g. Zhang et al, Am J Hum Genet 2015)? Can all the differences be explained by the surveying of the whole genome instead of whole exome? Are APOBEC mutations enriched in non-coding regions?
7. The authors carry out analyses of associations between mutational signatures and risk factors using a standard regression model. What about employing approaches that are more widely used in the epidemiology field, e.g. mixed models that would allow to distinguish between fixed and random effects? Furthermore, is it possible that multiple risk factors contribute to the same mutational signature? It may be beneficial to explore interaction effects or quantifying risk factor importance in the model. Also, signatures were dichotomised in the regression model: have the authors tried using different cut-offs for the presence/absence of a particular signature, or modelling on the continuous signature values?
8. Because most of the analyses compared countries individually rather than grouping them into high- and low-incidence, I am still left wondering for most of these results whether any differences between high and low incidence regions could have been captured: e.g. differences in positively selected genes (not tested for), differences in copy number changes, differences in APOBEC clonality (not tested for) etc.

Minor comments:
1. Was the tumour stage annotated for these samples and were differences in signature prevalence observed by cancer stage? This should also be accounted for in the regression analysis.
2. In several places, the authors reference the entire figure when they should be referencing a specific panel of the figure. Please specify e.g. Fig 4a/b/c when referring to the figure (not only Fig 4), to avoid the reader having to put extra effort to identify the specific panel you are referring to.

Reviewer #2:
Remarks to the Author:
Moody et al., present a mutational signature analysis of 552 ESCC tumors from eight countries across three continents and the UK. The studied populations have very different rates of ESCC incidence. The hypothesis tested was that regional mutagens might impact ESCC tumor mutation rates. Therefore, they sequenced 552 of these tumors and tested for signatures of single base substitution, indels, double base substitution, and structural variants. They found no major differences in the spectrum of mutations across populations. 6 COSMIC signatures accounted for 80% of all mutations. Two APOBEC signatures were detected in 90% of samples accounting for 25% of all mutations. Other signatures - SBS1 and SBS5/SBS40 were related to aging, while SBS18, believed to be caused by reactive oxygen species, accounted for 55% of mutations. Interestingly, they found a significant contribution of an HRD-associated signature, specifically SBS3, which they found in 8% of the patients. ID and DSB analysis led to other signatures with similar proportions across populations and mostly without any

known etiology. The structural variant analysis was mentioned but not discussed in sufficient depth. The authors then looked for associations of signatures with populations, risk factors, and germline mutations and found modest associations. SBS16, DBS4, and ID11 were more common in Brazilian and Japanese individuals. They reported associations between SBS16 and alcohol consumption and the presence of the rs671 SNP. They also found associations of SBS288J with opium use.

This work is comprehensive with importance for the global epidemiology of ESCC. As the authors wrote, the lack of a clear smoking gun in the form of dominant mutational signatures specific to regions with high ESCC incident rates points to non-DNA mutating agents as possible causes of these global variations in ESCC incident rates. I have a few comments on the methodology and suggestions for analyses that can strengthen the paper.

1) In Figure 2 the authors show de-novo extraction and then project them on COSMIC but used the COSMIC (PCAWG) etiology to explain the results. This is confusing since the authors published a paper (Alexandrov et al., 2020, Nature) where they used >2500 WGS and found the COSMIC signatures as building blocks. The novel signatures are decomposed and many of these are composed of multiple COSMIC signatures. After the authors did all of this work to extract de-novo signatures, they didn't use the SBS288 but rather the COSMIC/PCAWG signatures with regard to signature activity attribution per sample. On such a sizable cohort, I would not have expected such procedures to be necessary. Is there a reason the authors believe further decomposition to the COSMIC signatures is the more biologically correct procedure, and not that the signatures they have extracted best describe the ESCC rather than a pan-can average? Are the confounding technical factors in the cohort that cause signatures of the same etiology to get extracted in multiple versions, necessitating the need to project to COSMIC? If so what are the implications for the accuracy of attribution? I would like to see more explanations of the process in the methods and to expand the discussion section.

2) Many of the Brazilian individuals studied have Japanese ancestry. Did the authors correlate Brazilian & Asian (or even Japanese) ancestry with SBS16/DBS4/ID11 levels?

3) SBS16, based on https://cancer.sanger.ac.uk/cosmic/signatures/SBS/SBS16.tt, is associated with transcription-coupled DNA damage. Did the authors confirm the relevant strand asymmetry pattern, which is seen in Liver cancer, also in ESCC?

4) SBS288J, which is linked to opium use, is decomposed to APOBEC and SBS5 signatures. However, to me, it looks like the TCT> TGT signature. Can the authors run the association of this channel alone with opium?

5) BRCA1 and BRCA2 generate different signatures as the last author showed in Nik-Zanial et al. Nature, 2016. Can the authors look at the type of copy number? The authors should run HRD detect developed by Stratton and colleagues or CHORD (https://www.nature.com/articles/s41467-020-19406-4) to determine the type of HRD status either BRCA2 or BRCA1. In addition, can the authors add the following additional columns to Supplementary Table 18: biallelic status, HRD related signatures activity, HRDetect and CHORD HRD-type status?

6) Since this is a large dataset, the author should run a gene burden test of germline PTVs as was performed in PCAWG marker paper (https://www.nature.com/articles/s41586-020-1969-6). And germline+somatic PTV association test to discover possible new genes.

7) Do samples with high SBS288H harbor NTHL1 mutations? Do samples with high Sig10A harbor POLE mutations?

Reviewer #3:
Remarks to the Author:
This is a unique project aiming to explore putative novel environmental risk factors through the study of mutational signatures in esophageal squamous cell carcinoma (ESCC) tumors from regions with vastly different rates of this cancer. The team is uniquely set up to conduct this project through the strong collaboration between the Sanger/US teams with IARC with its strong ties to cancer research teams in many countries included in this large endeavor. The researchers should be applauded for taking on this very momentous and novel study. Bringing these diverse samples together is an enormous achievement; however, it is disappointing to see the limited set of lifestyle/epidemiological risk factors that have been included. Furthermore, the limited description of the included studies which seems mix of existing studies and newly recruited patients. It is not clear how comparable these different studies are as details are missing. This is an important limitation that needs to be addressed.

It seems surprising that the finding for opium exposure and the SBS288J does not get more attention. Is this because of the difficulties to replicate? Could additional samples from Iran be recruited or an animal study be done to add an independent validation?
It is disappointing that not more environmental risk factors were collected across the studies and that opium exposure was only available in Iran.
A major question is what findings are novel given the previous studies (such as ref 13-20) and how do these agree and disagree. This should be summarized in the discussion.
SBS5 and SBS40 combined accounted for 38% of the mutation burden, which is sizable. The etiology of SBS5 and SBS40 is unknown. Could this point to a novel exposure that has not been defined?
Why were Copy number profiles only investigated in a subset of cases.
Line 120: add the # of hypermutators.
To get an overall senses of the quality of the material used it will be important to add to the result section what material was used for DNA extraction, such as FFPE, fresh frozen, and for normal (surrounding normal tissue, blood, buccal,…). If a mix was used add #s.
The abstract should add that the signatures due to activity of the APOBEC (SBS2 and SBS13) were present in 88% and 91%. This information seems at least equally interesting as the information that these signatures account for 25% of the total mutation burden.
Line 145 following describes the predominant signatures and their presence and contribution to the mutational burden. It would be good if for each of these common occurring signatures both presence across the tumors and % contribution to the mutational burden can be described. Currently this is not consistently done which is confusing, particularly for the DBS.
Line 198-99: this sentence is missing the region where SBS16 is higher: "SBS16 mutation burden was significantly higher when compared to all other regions"
Line 255: describe the acronym TCN
Line 281: "…whereas the de novo signature is a composite…" Is the de novo signature SBS288I
Germline genetic associations: The authors only describe that the frequency of the germline deletion in APOBEC3 in ESCC cases from China and Japan was lower than expected in the general population, supporting a potential protective role in ESCC (0.28 vs 0.37); however, there is no mentioning of the other regions. What are the findings in those? Furthermore, the discussion (line 366 and following) is

not sufficient. What could be the reasons why the finding are in opposite direction from previous publications?
The discussion on the germline and somatic BRCA1/2 variants, and potential treatment options should point out the low frequency of these occurrences.
While the investigators provide more details on the prospective collection of cases no information on the selection of cases and assessment of biospecimens and risk factors for ongoing/retrospective case series has been provided. It is also not clear how many samples were collected prospectively vs through ongoing/retrospective case series. How do the authors define prospective and retrospective studies?
Each study included needs to be described in more detail, including method or recruitment, area and years of recruitment, inclusion, exclusion criteria,… Furthermore a table showing the clinical characteristics and epidemiological variables should be shown by study (no just by region).
It would be good to provide details about the concordance of the 28% of the cases that underwent at least 2 independent pathology examinations.
In the methods it is stated that DNA extraction was conducted centrally or locally. However, as IARC/WHO coordinated a centralized digital pathology review how was the tumor DNA extracted locally. Were digital images shared between local centers and coordinating center?
The sequencing read depth can differ substantially by genomic region therefore it is better to used exclusion cut off points based on x% covered at least x reads (such as 80% of reads at 30x) instead of overall coverage (which was set to below 30X for tumor or 15X for normal tissue).
50 Japanese ESCC tumors with paired adjacent esophagus were sequenced separately using a different lab and platform. How were these data integrated in the analyses and ensured that some of the findings in Japanese cases are not due to this potential batch effect?
Epidemiological risk factors were collected using different questionnaires from retrospective and prospective studies. The data harmonization needs to be described in detail to better understand potential limitations. Ideally the authors include the exact variables assessed in each study and describe how the variables were assessed in each study.
For germline genetic association testing the investigators only adjusted for sex, age of diagnosis and country of origin. The analysis should also adjust for genetic determined ancestry.

This reviewer might have missed it but details on how the studies were combined in the regression analysis (pooled vs meta-analysis) should be added and how region and the different (retrospective and prospective) studies were accounted for in the analyses.

Figure 1 includes a table of the distribution of the exposure variables that have been considered. Several exposures have sizable # of missing variables or very limited variability in some regions. This will substantially impact the power which should be pointed out as a limitation.

Extended Data Fig.1-3 and elsewhere: Would it be better to list regions by ESCC rates than alphabetically?

Author Rebuttal to Initial comments

Author comments:

Please note that in order to reduce the word count, several sections with non-significant results have been moved to a supplementary results section. These include copy number profiles, additional information on *de novo* signatures which could not be decomposed, rearrangement signatures and clustered mutations. All results in the supplementary section are non-significant and are not necessary to convey the principle findings of the study.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The authors present an overview of mutational and structural variant signatures inferred from 552 ESCC genomes which have been collected from multiple high and low incidence regions, and relate these patterns with epidemiological parameters in view of identifying potential new exogenous sources for the observed mutational signatures.

The study benefits from well curated data acquired through an epidemiological questionnaire (not extremely "extensive", but quantifying useful parameters). The authors have been very thorough regarding their approaches for inferring the mutational signatures: they test and compare multiple variant callers, and have also tried two different mutational signature methods, even though the latter is not discussed in great detail. This gives confidence that the mutational signatures inferred are as robust as one might expect them to be given the usual caveats. The statistics employed are appropriate for this type of study.

The study recapitulates epidemiological associations with mutational signatures in ESCC which had been reported in other studies, it highlights a role for DDR deficiencies in ESCC (with BRCA1/2 variants observed in a subset of cases) and it identifies a potential new signature which may be linked to opium consumption (however, not clearly distinguishable as stand-alone). Appropriate credit was given to previous work. Most of the results are summarised clearly, but the figures are not always the best suited to illustrate differences between groups (as specified in my comments below).

This is a study with robust methodology, but overall seems to lack the novelty that would be expected for a Nature Genetics publication. The claimed aspects of novelty are preliminary and rather speculative in nature.

Major comments and suggested improvements:
1. HDP was run in addition to SigProfiler to infer mutational signatures. While HDP results are briefly presented in the extended figures, they are difficult to compare to the ones from SigProfiler. It would be useful if the authors could generate some summary statistics or visualisations that highlight to what extent the results from the two methods agree.

Thank you for pointing out this limitation. In response, we investigated how similar the final results would be using either the SigProlifer of HDP *de novo* signatures as the starting point. HDP does not feature the capability to decompose into COSMIC reference signatures. Instead we have analyzed them in the same manner as the SigProfilerExtractor *de novo* signatures by first decomposing the *de novo* HDP signatures using the SigProfilerExtractor decomposition module and then attributing signature activities using MSA. Methods have been amended to include this. This data was then used to replace Extended Data Fig 2b so that it can now be directly compared to Fig3b. We have also added an additional panel (Extended Data Fig. 2c), which compares the average attribution from HDP and SigProfilerExtractor results to each of the signature groups plotted in the previous figure. This shows that methods agree for most groups, but there are some differences in the flat signatures (SBS3 and SBS5/SBS40) which is to be expected given the uncertainty in extracting and attributing such signatures.

2. The authors briefly mention a signature SBS288P of unknown aetiology which has also been reported in stomach cancers and is present in 5% of the samples. Could the authors clarify which countries this signature has been observed in and at what prevalence in each? Was this signature also captured using HDP?

Prevalence of SBS288P in each country has been added to the main text. This signature is also captured by HDP (Component 21) and the text has also been amended to include this. In order to reduce the word count and as no significant results are reported involving SBS288P, this section has been moved to supplementary results.

3. The authors investigated copy number profiles "in a subset of cases" (line 124). It is not clear why copy number variation could not be investigated in the entire cohort. The average copy number profiles were presented by country and indeed they look similar, but it is difficult to quantify from this visualisation whether any regions may differ significantly. Beyond the total number of segments altered, are there specific regions that are systematically changed (e.g. amplified or deleted) by country? Were there any cancer drivers specifically amplified or deleted only in certain countries?

Only cases with tumour purity >50% were included due to the influence of tumour purity on the algorithms used to determine copy number. The text has been amended to clarify this point. Whilst we thought it was useful to provide the high-level overview of the copy number changes we do not think this data set would provide reliable enough profiles to investigate differences in copy number drivers in detail. Previous studies such as PCAWG have utilized multiple callers to construct consensus CNV profiles, which was considered outside the scope of this study. Even if this was attempted, for approaches such as GISTIC which is similar in concept to dNdScv in using a statistical approach to identifying areas of recurrent amplification and deletion, it would be inappropriate to compare the regions due to very limited power in low incidence countries compared to the higher incidence locations. Given these limitations and the fact that no

significant results are reported, we have moved this section to the new Supplementary Results section.

4. The authors analyse the TP53 mutation spectrum and compare it between smokers/non-smokers and drinkers/non-drinkers. It was not clear to me whether the authors are looking here at mutations acquired only within the TP53 gene itself or at something else (TP53 mutated vs wild type cancers)? It would be useful if the authors clarified how they derive the spectrum in Figure 4a. Also, an enrichment of SBS16 mutations is reported in drinkers, but the plots in Figure 4 are not helpful in highlighting this. In fact, I do not think any of the panels in Figure 4 are particularly helpful in illustrating the claims the authors are making in the paragraph spanning lines 238-249. Could the authors find an alternative, more quantifiable visualisation in addition to the mutational spectra provided?

We have clarified in the text that the TP53 spectra is constructed from the TP53 SBS variants which were identified in the ESCC cases. We acknowledge the limitations pointed out regarding Fig.4, and have revised it accordingly. For the first panel we have added the plots for the APOBEC associated COSMIC reference signatures SBS2 and SBS13. We have enhanced Figure 4d, with arrows now indicating the SBS16 compatible contexts. In addition, two additional panels (Fig.4e and g) have been added to show the enrichment of mutations in SBS16 compatible contexts in drinkers, but comparable C>A mutations in smokers vs non-smokers.

5. The authors stress on the importance of the APOBEC mutational process in the development of ESCC. While it is clear this is likely to have a strong contribution, this has been reported before and it does not seem the clonality analysis in bulk samples of varying purity would be of high enough confidence to make any claims about the timing of APOBEC mutagenesis (even though it is useful to include this as a minor point). The non-significant results in Figure 5 would have fit better in the supplementary material, and do not justify it being a standalone figure by itself.

Thank you for this valid point. Fig. 5 has been moved to the extended data figures.

6. The authors speculate APOBEC mutagenesis might be a mandatory step for ESCC oncogenesis – however, there is not enough evidence to support that it acts with certainty in all cases, or its link to TP53 loss of function. From the figures presented, it was difficult to estimate the exact number of cases with an APOBEC signature, and the overall prevalence of this signature across all the cases in all cohorts (beyond country-level averages). How do the authors explain that other studies have only found evidence for APOBEC mutagenesis only in a subset of tumours (e.g. Zhang et al, Am J Hum Genet 2015)? Can all the differences be explained by the surveying of the whole genome instead of whole exome? Are APOBEC mutations enriched in non-coding regions?

Thank you for this point. The exact number of cases with APOBEC signatures is shown in Figure 2b, with prevalence stated in the text. There are multiple factors that likely explain why previous studies have detected APOBEC in a smaller subset of cases. We ran SBS288 signature extraction SigProfilerExtractor on our cases again, limiting to the exome to explore this further. This results in only 6 *de novo* signatures being extracted (as compared to the 20 extracted in the whole genome analysis). SBS2 and SBS13 were each present in 72% of the cases, compared to 88% and 91% of cases in the full analysis. This is still greater than some previous studies, in the paper mentioned (Zhang et al 2015), they identified an APOBEC signature present in 47% of cases, but this is likely explained by the other factors, such as the larger cohort size available for this study and the fact that methodologies for both extracting and attributing mutational signatures have improved greatly since some of the previous studies were published. In regard to whether APOBEC mutations are enriched in non-coding regions, as there is currently no reference set of SBS288 signatures we examined the *de novo* signatures SBS288A and SBS288B which closely resemble the COSMIC references signatures SBS13 and SBS2. The probability of mutations occurring in non-transcribed regions was 57% and 64% respectively, 22 and 18% respectively for transcribed regions, and 21% and 18% respectively for untranscribed regions. These values are similar to what one would expect overall in the genome, suggesting that there is little enrichment between different regions of the genome, although a more comprehensive analysis would be needed to confirm this.

7. The authors carry out analyses of associations between mutational signatures and risk factors using a standard regression model. What about employing approaches that are more widely used in the epidemiology field, e.g. mixed models that would allow to distinguish between fixed and random effects? Furthermore, is it possible that multiple risk factors contribute to the same mutational signature? It may be beneficial to explore interaction effects or quantifying risk factor importance in the model. Also, signatures were dichotomised in the regression model: have the authors tried using different cut-offs for the presence/absence of a particular signature, or modelling on the continuous signature values?

We appreciate the reviewer's well-founded comments with regards to associations of mutational signatures with risk factors. Mixed models were attempted but did not provide any benefit, as illustrated in the new Supplementary Table 21. For three risk factors and corresponding signatures, the base model 'Base_glm' is compared with a mixed model with a random intercept but no random effects (coded as 'Base_lmer_Mod'), as well as a mixed model with both random intercept and random effect for the exposure (coded as 'RE_lmer_Mod'). Notably, mixed models with a random intercept yielded very similar results in terms of significance, and marginally higher AIC. In addition, more complex models with random effects showed singular variance, suggesting that associations are essentially the same across different regions. The use of mixed models was not possible for signatures observed in a small fraction of the cases, such as SBS16, due to perfect or near-perfect separation of data, which is why a penalized approach (Firth method) was used instead in such cases. Investigation of the interaction effects was unfortunately

unfeasible in this dataset due to the lack of power. The continuous signature values as well as different cut-offs for the presence and absence were also tested, but did not show any meaningful difference in the outcomes. For these reasons, the authors chose the simplest approach of a logistic regression.

8. Because most of the analyses compared countries individually rather than grouping them into high- and low-incidence, I am still left wondering for most of these results whether any differences between high and low incidence regions could have been captured: e.g. differences in positively selected genes (not tested for), differences in copy number changes, differences in APOBEC clonality (not tested for) etc.

We appreciate this comment. In regard to showing differences in individual countries as opposed to low/high incidence, we felt it was not appropriate in all cases, for example, SBS16 enrichment is only observed in two out of three of the lower incidence regions (UK, Brazil and Japan). Similarly, we felt it would be inappropriate to assume that the factors driving high incidence regions would be the same, given that the incidence rate is highly variable even within the countries typically considered to be high incidence. Furthermore, this type of analysis was limited to due to the small number of cases available for the low incidence areas. Performing the dNdScv analysis for low and high incidence regions results in 11 significant genes for low incidence regions and 35 significant genes for high incidence. All 11 genes in the low incidence regions were also significant in the high incidence region. However, it is not possible to distinguish whether genes appearing in the high incidence group only are due to them only being selected for in these regions, or whether there is simply not enough power to detect in the low incidence region, especially considering that the frequency of mutations in any individual gene (with the notable exception of TP53) is 15% or less in our cohort. We also tested running the analysis in cohorts with >100 cases (Iran, China, and East Africa), with no differences observed. For APOBEC clonality however, this analysis is possible, and we have added a second panel to Extended Data Fig 8 (previously Fig. 5) which shows each country individually. No significant differences were found in any individual country (the UK could not be tested as only 2 cases fit the criteria) and the text has been amended to include this finding.

Minor comments:
1. Was the tumour stage annotated for these samples and were differences in signature prevalence observed by cancer stage? This should also be accounted for in the regression analysis.

Unfortunately, the tumour stage was annotated only for Asian population, making it impossible to account for in general regression analysis. However, most ESCC tumors are usually diagnosed at a late stage, particularly in African cohort. (https://pubmed.ncbi.nlm.nih.gov/27466161/)

2. In several places, the authors reference the entire figure when they should be referencing a

specific panel of the figure. Please specify e.g. Fig 4a/b/c when referring to the figure (not only Fig 4), to avoid the reader having to put extra effort to identify the specific panel you are referring to.

Apologies for this oversight, the text has been amended accordingly

Reviewer #2:
Remarks to the Author:
Moody et al., present a mutational signature analysis of 552 ESCC tumors from eight countries across three continents and the UK. The studied populations have very different rates of ESCC incidence. The hypothesis tested was that regional mutagens might impact ESCC tumor mutation rates. Therefore, they sequenced 552 of these tumors and tested for signatures of single base substitution, indels, double base substitution, and structural variants. They found no major differences in the spectrum of mutations across populations. 6 COSMIC signatures accounted for 80% of all mutations. Two APOBEC signatures were detected in 90% of samples accounting for 25% of all mutations. Other signatures - SBS1 and SBS5/SBS40 were related to aging, while SBS18, believed to be caused by reactive oxygen species, accounted for 55% of mutations. Interestingly, they found a significant contribution of an HRD-associated signature, specifically SBS3, which they found in 8% of the patients. ID and DSB analysis led to other signatures with similar proportions across populations and mostly without any known etiology. The structural variant analysis was mentioned but not discussed in sufficient depth. The authors then looked for associations of signatures with populations, risk factors, and germline mutations and found modest associations. SBS16, DBS4, and ID11 were more common in Brazilian and Japanese individuals. They reported associations between SBS16 and alcohol consumption and the presence of the rs671 SNP. They also found associations of SBS288J with opium use.

This work is comprehensive with importance for the global epidemiology of ESCC. As the authors wrote, the lack of a clear smoking gun in the form of dominant mutational signatures specific to regions with high ESCC incident rates points to non-DNA mutating agents as possible causes of these global variations in ESCC incident rates. I have a few comments on the methodology and suggestions for analyses that can strengthen the paper.

1) In Figure 2 the authors show de-novo extraction and then project them on COSMIC but used the COSMIC (PCAWG) etiology to explain the results. This is confusing since the authors published a paper (Alexandrov et al., 2020, Nature) where they used >2500 WGS and found the COSMIC signatures as building blocks. The novel signatures are decomposed and many of these are composed of multiple COSMIC signatures. After the authors did all of this work to extract de-novo signatures, they didn't use the SBS288 but rather the COSMIC/PCAWG signatures with regard to signature activity attribution per sample. On such a sizable cohort, I would not have expected such procedures to be necessary. Is there a reason the authors believe further decomposition to the COSMIC signatures is the more biologically correct procedure, and not that

the signatures they have extracted best describe the ESCC rather than a pan-can average? Are the confounding technical factors in the cohort that cause signatures of the same etiology to get extracted in multiple versions, necessitating the need to project to COSMIC? If so what are the implications for the accuracy of attribution? I would like to see more explanations of the process in the methods and to expand the discussion section.

Thank you for pointing this out. Firstly, extracting the SBS288 *de novo* signatures is a necessary step in our methodology, as only the COSMIC reference signatures identified via decomposition of the *de novo* signatures are included in the panel used for attribution. This improves the attribution by removing reference signatures which are not present in the cohort prior to attribution. This point has been clarified in the methods. Secondly, the *de novo* signatures are composites of multiple signatures, whereas the COSMIC signatures are more likely to represent individual biological mechanisms, which provide the better option to identify associations with epidemiological risk factors. Lastly whilst we used both *de novo* and COSMIC signatures where possible in regression analysis to increase confidence in the findings, we felt that presenting the results in COSMIC format was more accessible to readers, as whilst the *de novo* signatures are unique to this study, some readers will have an awareness of COSMIC reference signatures and their mechanisms. The discussion has been expanded to include a statement on mutational signature methodology.

2) Many of the Brazilian individuals studied have Japanese ancestry. Did the authors correlate Brazilian & Asian (or even Japanese) ancestry with SBS16/DBS4/ID11 levels?

According to the newly added principal components analysis performed to determine genetic ancestry, now described in the Methods section, there were no Brazilian individuals with Japanese ancestry in our cohort (see Extended Data Fig. 9 with principal components).

3) SBS16, based on https://cancer.sanger.ac.uk/cosmic/signatures/SBS/SBS16.tt, is associated with transcription-coupled DNA damage. Did the authors confirm the relevant strand asymmetry pattern, which is seen in Liver cancer, also in ESCC

The strong transcription bias previous observed on T>C mutations is present in the *de novo* signature SBS288Q (which decomposes into SBS16), as shown in Figure 2c.

4) SBS288J, which is linked to opium use, is decomposed to APOBEC and SBS5 signatures. However, to me, it looks like the TCT> TGT signature. Can the authors run the association of this channel alone with opium?

The SBS288J which decomposes into APOBEC and SBS5 is the signature obtained from the extraction including all cases and does not associate with opium use. It is SBS288J from the Iranian regional extraction that associates with opium use, this signature decomposes into SBS1,

SBS5, SBS12 and SBS25. We have renamed all regional signatures to include the region (in this case SBS288J_Iran) to make this clearer. The predominant channel in this signature is T>C, and running the association of this channel alone with opium does in fact yield a positive result, however, we chose to present the results of the whole signature, in keeping with the other results presented.

5) BRCA1 and BRCA2 generate different signatures as the last author showed in Nik-Zanial et al. Nature, 2016. Can the authors look at the type of copy number? The authors should run HRD detect developed by Stratton and colleagues or CHORD (https://www.nature.com/articles/s41467-020-19406-4 [nature.com]) to determine the type of HRD status either BRCA2 or BRCA1. In addition, can the authors add the following additional columns to Supplementary Table 18: biallelic status, HRD related signatures activity, HRDetect and CHORD HRD-type status?

Thank you for this interesting comment. No associations were observed in relation to the structural rearrangement signatures, in contrast to previous studies. This point has been added to the text in the Supplementary Results. We ran CHORD analysis which was designed for pan-cancer HRD detection, however many of the cases with high confidence pathogenic variants in BRCA1/2 were not detected as being HR deficient. In addition, all cases which were HR deficient were called as type 2, even in cases with high confidence BRCA1 mutations. This suggests that the tool is not performing optimally in this tumour type, or may not deal well with samples of varying purity such as those in this cohort. We have added biallelic status, CHORD status, and HR related signature activity to Supplementary Table 18 as requested, however as some cases were of insufficient purity to determine a copy number profile and as we do not have data on promoter methylation, there are some cases which remain unconfirmed in regards to biallelic status.

6) Since this is a large dataset, the author should run a gene burden test of germline PTVs as was performed in PCAWG marker paper (https://www.nature.com/articles/s41586-020-1969-6 [nature.com]). And germline+somatic PTV association test to discover possible new genes.

We appreciate the reviewer's comment regarding the potential for gene discovery in this context. However, while sufficient for the hypothesis we explore in this manuscript, we note that sample size remains relatively modest for genome wide discovery approaches, particularly given the heterogeneous nature of ethnicities included here. As such, we feel that this analysis is outside the scope of the current manuscript. The total sample size of the Mutographs project is envisaged to be in the order of 5,000 cancer and normal genomes, and we do anticipate this type of analysis when the entire dataset is completed.

7) Do samples with high SBS288H harbor NTHL1 mutations? Do samples with high Sig10A

harbor POLE mutations?

Only one NTHL1 variant was detected in a case with moderate SBS288H attribution, although NTHL1 was not included in the panel of drivers screened for, checks on this variant reveal it is of unknown significance. Similarly, the single case with high attribution to SBS10a contained a POLE variant of unknown significance. These associations have been better described elsewhere and so have not been included in this manuscript.

Reviewer #3:
Remarks to the Author:
This is a unique project aiming to explore putative novel environmental risk factors through the study of mutational signatures in esophageal squamous cell carcinoma (ESCC) tumors from regions with vastly different rates of this cancer. The team is uniquely set up to conduct this project through the strong collaboration between the Sanger/US teams with IARC with its strong ties to cancer research teams in many countries included in this large endeavor. The researchers should be applauded for taking on this very momentous and novel study. Bringing these diverse samples together is an enormous achievement; however, it is disappointing to see the limited set of lifestyle/epidemiological risk factors that have been included. Furthermore, the limited description of the included studies which seems mix of existing studies and newly recruited patients. It is not clear how comparable these different studies are as details are missing. This is an important limitation that needs to be addressed.

It seems surprising that the finding for opium exposure and the SBS288J does not get more attention. Is this because of the difficulties to replicate? Could additional samples from Iran be recruited or an animal study be done to add an independent validation? It is disappointing that not more environmental risk factors were collected across the studies and that opium exposure was only available in Iran.

The opium signature is difficult to extract for several reasons, primarily because it does not produce a strong signal relative to the total mutation burden which can be easily detected against the other mutational signatures present, particularly due to the presence of other T>C rich signatures in our cohort. For this reason, model systems likely represent the best option for cleanly identifying this signature however it was not possible to obtain this data for the present study. While opium consumption is relatively common in Iran, it is not frequent enough to allow meaningful analysis in the other populations examined in this study.

A major question is what findings are novel given the previous studies (such as ref 13-20) and how do these agree and disagree. This should be summarized in the discussion.

Thank you for this observation, the discussion has been amended in the appropriate sections accordingly

SBS5 and SBS40 combined accounted for 38% of the mutation burden, which is sizable. The etiology of SBS5 and SBS40 is unknown. Could this point to a novel exposure that has not been defined?

What we know about these signatures currently would suggest this is not the case, although we cannot completely exclude the possibility that there is an unknown exposure with a very similar mutation profile to these signatures. SBS5 exhibits clock like properties, in most cases the number of mutations is associated with age. It is found in the vast majority of tissues, both normal and tumour, albeit with different mutation rates. Rather than being a novel exposure itself, SBS5 is likely to be an endogenous process which can be influenced by environmental exposures, such as been found for tobacco exposure. SBS40 is also found in both tumour and normal tissues, and associates with age in some cancer types. SBS40 is prominent in kidney tissue, and it is hoped that the Mutographs cohort of renal cell carcinomas may further increase understanding of this mutational signature.

Why were Copy number profiles only investigated in a subset of cases.

Copy number profiles were only investigated in cases with tumour purity >50% due to the influence of tumour purity on the algorithms used to determine copy number. The text has been amended to clarify this point. Please note that copy number profiles have been moved to a supplementary results section.

Line 120: add the # of hypermutators.

The number of hypermutators has been added to the text as requested

To get an overall senses of the quality of the material used it will be important to add to the result section what material was used for DNA extraction, such as FFPE, fresh frozen, and for normal (surrounding normal tissue, blood, buccal,…). If a mix was used add #s.

Apologies for this oversight, this information has been added to the methods section.

The abstract should add that the signatures due to activity of the APOBEC (SBS2 and SBS13) were present in 88% and 91%. This information seems at least equally interesting as the information that these signatures account for 25% of the total mutation burden.

The abstract has been amended accordingly

Line 145 following describes the predominant signatures and their presence and contribution to the mutational burden. It would be good if for each of these common occurring signatures both presence across the tumors and % contribution to the mutational burden can be described. Currently this is not consistently done which is confusing, particularly for the DBS.

The text has been amended accordingly. The number of cases positive for an individual signature can also be found in the TMB plots provided in Fig2 and Extended Data Figs 3&4.

Line 198-99: this sentence is missing the region where SBS16 is higher: "SBS16 mutation burden was significantly higher when compared to all other regions"

Apologies, the text has been revised to make it clear that this refers to Brazil and Japan.

Line 255: describe the acronym TCN

An explanation of this acronym has been added to the text. In addition, in the methods we have clarified that only C>T and C>G mutations at TCN contexts were used in this analysis, due to substantial overlap with COSMIC signature 18 in C>A TCN mutations.

Line 281: "…whereas the de novo signature is a composite…" Is the de novo signature SBS288I

Apologies for any confusion, the text has been amended to make this clearer.

Germline genetic associations: The authors only describe that the frequency of the germline deletion in APOBEC3 in ESCC cases from China and Japan was lower than expected in the general population, supporting a potential protective role in ESCC (0.28 vs 0.37); however, there is no mentioning of the other regions. What are the findings in those? Furthermore, the discussion (line 366 and following) is not sufficient. What could be the reasons why the finding are in opposite direction from previous publications?

Apologies for this error, this is in fact the frequency of the rs12628403 variant, acting as a proxy to the deletion (corrected in text). The lack of European cases in the Mutographs ESCC cohort (7 cases from the UK) made any comparison with the general population unfeasible. For African cases, the frequency of rs12628403 variant did not differ from that of the general population in Africa (≈ 0.01), yet this variant is rare and understudied in these populations. No comparison was possible for Iran and Brazil due to the lack of publicly available frequencies in general population. The text has been amended to reflect this. It is not clear why the findings are different to those reported previously, we have now included this in the discussion and highlighted that further investigation will be required.

The discussion on the germline and somatic BRCA1/2 variants, and potential treatment options should point out the low frequency of these occurrences.

The discussion has been amended accordingly

While the investigators provide more details on the prospective collection of cases no information on the selection of cases and assessment of biospecimens and risk factors for ongoing/retrospective case series has been provided. It is also not clear how many samples were collected prospectively vs through ongoing/retrospective case series. How do the authors define prospective and retrospective studies? Each study included needs to be described in more detail, including method or recruitment, area and years of recruitment, inclusion, exclusion criteria, Furthermore a table showing the clinical characteristics and epidemiological variables should be shown by study (no just by region).

The revised manuscript now contains a table (Supplementary Table 19) which provides information on each individual study. With regards to prospective/retrospective cases, we apologize for the confusion: in fact, all ESCC cases presented in this study were retrospective. Initially, the methodology section was describing the general protocol used in Mutographs for case collection (both prospective and retrospective), which would be the case for subsequent cancer types, but not ESCC where only retrospective studies were analyzed. The methods have now been updated accordingly.

It would be good to provide details about the concordance of the 28% of the cases that underwent at least 2 independent pathology examinations.

This information has been added to the methods

In the methods it is stated that DNA extraction was conducted centrally or locally. However, as IARC/WHO coordinated a centralized digital pathology review how was the tumor DNA extracted locally. Were digital images shared between local centers and coordinating center?

Apologies for the confusion. Cases from Japan and the UK completed pathology evaluation locally under the same criteria used for the cases undergoing review at IARC. The methods have been corrected to reflect this.

The sequencing read depth can differ substantially by genomic region therefore it is better to used exclusion cut off points based on x% covered at least x reads (such as 80% of reads at 30x) instead of overall coverage (which was set to below 30X for tumor or 15X for normal tissue).

We used a mean coverage threshold of 30X for tumors, which was the threshold used in the PCAWG study, as this was sufficient to remove cases which had an obvious reduction in coverage relative to the rest of the cohort. We acknowledge the additional value of an evenness of coverage metric, and we did consider implementing the full PCAWG QC pipeline (https://www.nature.com/articles/s41467-020-18688-y) which includes evenness of coverage metrics. However, they noted that large copy number events could mislead such calculations. Given that ESCC is a structurally complex tumor, we had concerns that such an approach would be too stringent for our study and so did not implement them into our QC process. We examined our cohort using the threshold suggested by the reviewer, which reveals that 39 of the 552 cases (7%) would fail this threshold. Unfortunately, this includes 14/37 (38%) of the Japanese cases through which a number of significant observations were made and which would be weakened if the set was reduced in this manner. To assess whether the cases that did not pass the evenness metric might have been unduly influential in generating these observations, we tested whether there were any significant differences in relative signature attribution in Japanese cases passing this threshold vs those falling below it, with no significant differences detected. The same was true for China and Iran, although in these cohorts the proportion of cases failing was much smaller (both 5%), and the remaining countries could not be tested due to the low number of cases falling below the threshold. This suggests that including the small number of cases falling below the suggested threshold is unlikely to have resulted in any false positive results. On balance, therefore, we believe that excluding these cases, (particularly those from the already limited low incidence regions) would be more detrimental to the study than including them and would therefore prefer, if possible, to retain them.

50 Japanese ESCC tumors with paired adjacent esophagus were sequenced separately using a different lab and platform. How were these data integrated in the analyses and ensured that some of the findings in Japanese cases are not due to this potential batch effect?

Thank you for this valid technical point. The sequencing data received was converted back to FASTQ format and then mapped and analyzed using the standard analysis pipeline. The methods section has been amended to clarify this. Following consensus variant calling, no batch effects were observed in the Japanese samples, aside from the coverage discussed above. The findings in Japanese cases are backed up either by previous studies or with consistent results in other cohorts.

Epidemiological risk factors were collected using different questionnaires from retrospective and prospective studies. The data harmonization needs to be described in detail to better understand potential limitations. Ideally the authors include the exact variables assessed in each study and describe how the variables were assessed in each study.

Further details on data harmonization have been added to the methods section, with an accompanying Supplementary Table (Supplementary Table 20).

For germline genetic association testing the investigators only adjusted for sex, age of diagnosis and country of origin. The analysis should also adjust for genetic determined ancestry.

The authors appreciate this suggestion, the analysis adjustments have been made accordingly – as now outlined in the methods section.

This reviewer might have missed it but details on how the studies were combined in the regression analysis (pooled vs meta-analysis) should be added and how region and the different (retrospective and prospective) studies were accounted for in the analyses.

All our cases were pooled in regression analysis, and only retrospective studies were included in the analysis, as have now been corrected in the methods section. All retrospective studies were accounted for by harmonization procedure, as also outlined in the Methods section as well as the Supplementary Table 20. Different regions were included as covariates in regressions with epidemiological risk factors, as described in the Methods section.

Figure 1 includes a table of the distribution of the exposure variables that have been considered. Several exposures have sizable # of missing variables or very limited variability in some regions. This will substantially impact the power which should be pointed out as a limitation.

It is true that due to the differences in questionnaires of retrospective studies, some exposure variables were missing in particular cohorts. Notably, the variables used in data harmonization were not available in Brazil, Japan and the UK. This does impact the power to detect associations, which is now emphasized in the discussion section. The harmonization of available variables and its limitations are now described in detail in Supplementary Table 20.

Extended Data Fig.1-3 and elsewhere: Would it be better to list regions by ESCC rates than alphabetically?

Thank you for this suggestion, we have revised the order of countries in the figures accordingly

Decision Letter, first revision:

Our ref: NG-A56490R

22nd Apr 2021

Dear Mike,

How are you?

Thank you for submitting your revised manuscript "Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence" (NG-A56490R). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics. Please pass on my congratulations to Sarah and the rest of the team.

Have a lovely weekend.

Sincerely,

Safia Danovi
Editor
Nature Genetics

Reviewer #1 (Remarks to the Author):

The authors have addressed all my comments in a satisfactory manner. I recommend acceptance of the manuscript.

Reviewer #2 (Remarks to the Author):

The authors addressed most of my concerns in a satisfactory manner.

However, in the discussion section, it is not obvious to me how the authors addressed the question of differences in the signature results compared with their previous Nature 2020 Istudy.

This is an important point, as readers would like to put this study in the context of the previously published work. Most of the readers, including myself, using the authors' COSMIC database as a reference. But in this study, they obtained signatures that are a mixture of COSMIC.v3 signatures. So, the authors should spend a bit of time explaining the lack of concordance. I cannot see what is unique in the mutational landscape of esophageal squamous cell carcinoma that leads to discordance with their recent Nature paper.

Reviewer #3 (Remarks to the Author):

Thank you for the detailed response, however, describing
The authors need to delete the description pf the prospective cohort recruitment as only retrospective samples were included. This section needs to be replaced with a detailed description of how each cohort included were selected.
The authors need to discuss the limitations of using previously collected cohorts that were not collected with a standardized protocol and many years apart.
The potential negative impact that the differences in collection periods, such as China 1997 to 2005 and many other countries ~2015-2019, need to be discussed.

Author Rebuttal, first revision:

Reviewer #1:
Remarks to the Author:
The authors have addressed all my comments in a satisfactory manner. I recommend acceptance of the manuscript.

We thank reviewer 1 for their comments on our manuscript.

Reviewer #2:
Remarks to the Author:
The authors addressed most of my concerns in a satisfactory manner.

However, in the discussion section, it is not obvious to me how the authors addressed the question of differences in the signature results compared with their previous Nature 2020 Istudy.

This is an important point, as readers would like to put this study in the context of the previously published work. Most of the readers, including myself, using the authors' COSMIC database as a reference. But in this study, they obtained signatures that are a mixture of COSMIC.v3 signatures. So, the authors should spend a bit of time explaining the lack of concordance. I cannot see what is unique in the mutational landscape of esophageal squamous cell carcinoma

that leads to discordance with their recent Nature paper.

We thank the reviewer for their comments. The purpose of the Nature 2020 study was to provide a reference set of mutational signatures present in human cancers and to do so used a much larger cohort of almost 25,000 tumors (including whole genomes and exomes) from a wide variety of tissues displaying variability in the combinations of signatures present. The COSMIC reference signatures, many of which derive from the Nature 2020 study, therefore represent the current most comprehensive catalog of individual mutational processes.

In comparison, this study contained a smaller cohort from a single tissue, with very limited variability in the combinations of mutational signatures present. Due to these characteristics, we would not expect to be able to replicate the results from the Nature 2020 study, with most signatures extracted as combinations of multiple COSMIC signatures to varying extents, due to the fact that smaller number of samples may not allow complete separation of individual mutational signatures. For example, SBS2 and SBS13, attributed to the APOBEC deaminases and generally very strongly correlated with one another, have been separated only when an examined dataset contained samples that exhibited only one of the signatures (namely, clean SBS2 in UNG deficient samples). While such samples are rare, they were included in the Nature 2020 study allowing the complete separation of SBS2 and SBS13. No such samples are present in the current dataset and, as such, SBS2 and SBS13 do not completely separate during the *de novo* extraction. The decomposition step is therefore used to determine which *de novo* signatures can be explained by the known catalog of COSMIC mutational signatures, and which are previously unidentified mutational signatures. The methods have been amended to address this comment.

Reviewer #3:
Remarks to the Author:
Thank you for the detailed response, however, describing
The authors need to delete the description pf the prospective cohort recruitment as only retrospective samples were included. This section needs to be replaced with a detailed description of how each cohort included were selected.
The authors need to discuss the limitations of using previously collected cohorts that were not collected with a standardized protocol and many years apart.
The potential negative impact that the differences in collection periods, such as China 1997 to 2005 and many other countries ~2015-2019, need to be discussed.

We thank the reviewer for their comments. The methods sections have been revised accordingly to address these comments and draw attention to Supplementary Tables 19 and 20 which provide details of the cohorts included and the details/limitations of data harmonization.

Final Decision Letter:

In reply please quote: NG-A56490R1 Stratton

28th Jul 2021

Dear Mike,

Two accepts in a day! I don't know about you, but this is certainly a first for me..

I am delighted to say that your manuscript "Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence" has been accepted for publication in an upcoming issue of Nature Genetics.

Prior to setting your manuscript, we may make minor changes to enhance the lucidity of the text and with reference to our house style. We therefore ask that you examine the proofs most carefully to ensure that we have not inadvertently altered the sense of your text in any way.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A56490R1) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. <a

href="https://www.springernature.com/gp/open-research/transformative-journals"> Find out more about Transformative Journals

<B>Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of

your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A56490R1). Further information can be found at https://www.nature.com/nprot/.

Please extend my congratulations to Sarah and the rest of your team. I hope we get a chance to work together again soon.

With best wishes,

Safia

Safia Danovi
Senior Editor
Nature Genetics