nature portfolio

Peer Review File

Base-excision repair pathway shapes 5-methylcytosine deamination signatures in pan-cancer genomes

Corresponding Author: Dr Marc-Henri Stern

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors around Bortolini Silveira investigate cancer signatures, focussing on mutations arising from spontaneous Cytosine deamination, which leads to C>T transitions if not repaired. The two DNA glycosylases MBD4 and TDG are of special interest as they are highly efficient in detecting and processing G:T mismatches. While an association of cancer and mutated TDG was only reported in very very few cases, MBD4 defects in cancers are a well-known phenomenon and also previously investigated and reported by the authors.
In the first half (Figures 1 and 2), the authors perform bioinformatical analyses on a subset of single-base substitution patterns (SBS) showing a large contribution of C>T transition, which are derived from cancer whole-genome sequencing data of own and “public” origin. Many correlations are reported based on which the authors support (SBS1, SBS96) or disfavour (SBS95, SBS105, SBSnovel) an association of the C>T transition containing SBS and MBD4 defects. A brief but interesting excursion into CpA methylation and corresponding CpA>TpA mutations in UVM is undertaken, reporting a non-negligible amount such mutations in dependence of MBD4.
In Figure 3, the authors perform experiments in non-related (HAP1) cells to examine the genomic distribution of MBD4 and its relation with several epigenetic characteristics, such as chromatin accessibility and histone methylation. Contrary to the author’s expectations MBD4 is abundant in active chromatin regions such as “unmethylated” promoters. Switching back to cancer cells and public annotation-based sites of chromatin activity, they report that protection against C>T transitions is more efficient in transcriptionally active regions. A protective effect in early-replicating areas is also reported.
Finally the authors perform a long-term cultivation experiment to assess the impact of MBD4 and or TDG depletion in HAP1 cells. They found that MBD4 is the dominant enzyme in protecting against C>T transitions, with a small contribution of TDG, but only in the absence of MBD4. The replication-dependent bias seems to be connected to MBD4.
All in all, the authors provide data that corroborates previous observations and bring some new aspects into the picture.
Major comments:
MBD4 has already been reported several times to be a major contributor in the protection against C>T substitutions and validated in different cancers (especially in SBS1 and SBS96). The novelty factor of this statement is therefore rather low.
Finding two (to three) more signatures that could be de novo associated to MBD4 is rather interesting, the authors however reject this possibility based on rather superficial and correlative analyses. For example, a (not so big) strand-bias in SBS95 and SBSnovel, compared to no strand-bias in SBS1 and SBS96, is not enough ground to discard the possibility of MBD4-dependence, especially since no (mechanistic) analyses concerning genomic region, transcriptional activity, or actually replication timing etc. are performed.
While, I agree that the pattern of SBS105 does not (solely) link to 5mC deamination, as it is rather clear given the additional enrichment of other than C>T transitions (Fig 1a). The non-linear but quite distinct correlation between mCpG levels and CpG>NpG mutations, could still point to a mC-dependent mechanism, potentially involving dynamic mC oxidation by TETs and finally TDG-initiated BER. A clear exclusion of mC-dependence, as also mentioned in the discussion, is not appropriate.
Based on this, the authors should either provide more rigorous analyses/experiments or need to adapt their rather definitive phrasing (“SBS1 (age) and SBS96 (MBD4 deficiency) are the sole 5mC deamination signatures”) accordingly.

In general, the consistency of analyses performed and conclusions drawn is sometimes questionable. The manuscript would gain a lot of trust as well as strength if more analyses would be performed around the observed patterns and phrasing would be adjusted. E.g.

- While the effect of increased CpA>TpA mutations in UVM cells is quite distinct and interesting, the authors should provide

more structured comparisons. mCpA levels (Fig 2f) should, if available, also be compared between cancer tissues (analogous to the comparison in Fig2b), if possible stratified per cancer type, especially for the HGG/Oligodendrocyte-like tissue harbouring higher levels of mCpA.
-Although high methylation levels don't necessarily mean stable mC levels (https://doi.org/10.1038/s41467-020-16354-x). The hypothesis of more stable, and therefore more prone to deamination, mCpA in OVM cells due to lower turn-over by TETs (and TDG) is intriguing. It would however be more telling if the absolute expression values, as depicted in Fig 2g, would i) be separated into MBD4def and MBD4wt UVMs and ii) also be put into relation with TET expression levels in the corresponding healthy tissue to asses this potential causation. Also, this should be related to the observation that C>T transitions generally happen at highly methylated, i.e. potentially more stably methylated, sites (Supp Fig 10).
- To gain deeper information about "how", rather than just "that" MBD4 is such a frequent cancer driver, I suggest to translate expression analyses to the data in Figure 2h, where i) transcriptional levels of these genes could be correlated to the mutation rate (are these genes generally highly expressed in the corresponding UVM types) and ii), more importantly, whether these sites, where mutations are observed, show a high mC level or not, akin to what was performed by Robertson et al. (REF55), although not regarding the state of MBD4 or CpA. Potential differences or similarities between effects at CpA and CpG sites could be highlighted.
- Comparisons between the replication timing and examined chromatin marks (Fig. 3c) should be made, to estimate the connection of replication timing, accessibility, methylation state and finally mutation rate.
- It is not entirely surprising that MBD4 is binding not only to methylated promoters, as first, it is able to bind carboxylcytosine (10.1093/nar/gks714), which is not detected in a bisulfite sequencing, and second, its glycosylase activity is pretty much independent from the methylation state in the vicinity (https://doi.org/10.1128/MCB.00588-08). Even more so, it was already published that MBD4 frequently binds to promoters that, not only show mC abundance but several marks of active transcription (https://doi.org/10.1016/j.cell.2013.03.011). Furthermore, the analysis the authors use does not provide the resolution (with figure 4c or Supp 8) to deny that MBD4 is actually binding to localized mCpGs in an otherwise low-methylated promoter. Given the previous observations, it would be interesting to see the mCpA levels in HAP1 cells and a corresponding graph of mCpA levels in Fig 3c.

Unfortunately, the documentation of bioinformatical analyses and datasets generated, as well as information/visualization of quality control is insufficient (public datasets exempt):
- TDG and MBD4 CUT&Run datasets are not findable/accessible, and major information about the cell lines generated is missing. The "validation" immunofluorescence in Supp Fig. 11 has to be mentioned earlier, and obvious unspecificities cannot be neglected. E.g. uncropped western of potential degradation fragments in N-terminally tagged MBD4 could explain such a high number of peaks reported in Fig 3a. No information about number of replicates, sequencing depth, peak calling parameters (e.g. FDR and fold enrichment) are mentioned. Coverage tracks with higher magnification than Supp Fig 8. should be provided and/or a signal vs. Noise estimation on in the TDG experiment, as all available antibodies struggle with unspecificities.
-Apart from the cropped western blot in Fig 4a, no information about the newly established cell lines are given. This should at least include a display of verification PCRs on DNA and cDNA as well as duplication times of these clones (optimally monitored over the 120 day window of cultivation).
- No detailed information about the WGS sequencing is given for these cells. How many replicates per clone? How many reads, how high was the coverage? Some of those details are given for the patient-derived samples but it is not visible whether this is applicable to the HAP1 cells.
- The authors explanation on how the replication timing data was employed is very limited. I assume their mentioning of "pan-tissue" replication timing refers to regions which display similar RT among different tissue types. The authors need to elaborate on this and discuss how many sites they examined, and whether the examined tissues are part of that collection.
- In the analysis of DNA methylation blocks, the authors should provide the following information to get a better understanding of the data: the number of blocks retained for the analysis, the number of promotors /regulatory elements covered by these blocks and the (average) amount of CpGs in the blocks covered.

Minor comments:

In Fig 1a, the authors should label the x-axis with an exemplary "sequence context" and might highlight the most frequently enriched C>T transitions, as done in Fig 2b.
Although written in the text, it would be handy to have a (table-like) overview of how many samples/patients are represented in the individual SBS categories, and how many of those carry which (known) mutation. Otherwise it is difficult to estimate the extent of observed results.
Fig2. While the brain has provided multiple and early evidence for CpA methylation, the references for non-CpG DNA methylation are a little dated and, therefore the statement: "CpA methylation primarily accumulates in differentiated tissues of neural origin by DNMT3A-mediated de novo methylation42,43,44" should be adapted to incorporate more (recent) observations about non-CpG methylation, e.g. in stem cells, spleen, etc.

The authors mention that they "observed a similarly strong TDG binding in active
TSSs and active enhancers (Fig. 3c; Supplementary Fig. 8)", which is not an appropriate statement given the mentioned two figures. While an enrichment of TDG and MBD4 at TSS is visible, the signal(intensities) are not quantifiable/comparable through heatmaps, and the binding strength can most certainly not be assessed by ChIP/CUT&Run assays. A comparison of binding frequency can only be estimated by densitiy plots (integrated from the heatmaps), but actual quantitative comparison between MBD4 and TDG binding would require multiple level of normalization, including Antibody and chromatin spike-in methods, aside from other biochemical experiments for binding strength.

Fig3e: The authors mention a relative mutation rate. Relative to what? Also, the legend hints to two different significance

values (**<0.01 and ****<0.0001) while in the graph both comparisons show ****.
The authors mention that “[...] was significantly more pronounced in SBS1 MBD4wt cases than in SBS96 MBD4def cases (ratio of repressed mutation rate by active mutation rate: 3.77±1.20 [SBS1] vs 1.56±0.30 [SBS96]; Wilcoxon t-test P = 1.27e-11)”.
They compare two different signatures in two different genotypes. I believe that the same signatures should be compared over the genotypes, and the data should be visualized.

While sometimes indicated in the text, statistical indicators are missing for the whole Fig 4.

Reference 27 is corrupted

Reviewer #2

(Remarks to the Author)
In this study, the authors seek to validate the etiology of rare COSMIC mutation signatures using a wide cohort of tumor data sets spanning over 12000 genomes, specifically focusing on the 5-methylcytosine deamination-based signatures. 5-mC deamination in CG contexts generates C>T transitions, which are among the most abundant mutations in the mutational spectra of a majority of tumors; however, the relationship between 5-mc deamination and C>T centric SBS signatures is poorly understood. The authors observed a striking correlation between the prevalence of C>T changes and the expression of DNA glycosylases, particularly MBD4. A subset of MBD4 deficient tumors had a unique remarkable enrichment of the rare SBS96 signature, which previously had a somewhat confounding etiology. Further, the authors have done an adequate job of combining genome wide methylation and histone modification data with signature activities to understand the distribution of C>T mutations across the genome. Finally, the authors were able to recapitulate the contribution of glycosylases to the observed signatures using HAP-1 cell lines with knockouts for both the major 5mC-deamination glycosylases MBD4 and TDG.

Minor comments:

1) The contribution of MBD4 to CG>TG burden in UVMs is quite interesting. Have the authors considered a correlation between with C>T mutations generated from UV radiation, and see whether or not there is a significant overlap? In line with this, what proportion of the observed C>T mutations in MBD4wt and deficient tumors CC>TT dinucleotide changes
2) It was also perhaps surprising that there was no overlap between the oncogenic driver mutations between any of the MBD4def tumors. While tumor-specific selection can explain this to some extent, I’m curious if there are additional factors. Can you stratify the MBD4wt v MBD4def signatures by age across various tumor types? Perhaps there is some clock-like component to mutations observed in the absence of MBD4.
3) The choice of using diploidized HAP1 cells for clonal expansion experiments isn’t adequately explained. Please elaborate in the methods section.

Reviewer #3

(Remarks to the Author)
This paper titled “Base-excision repair pathway shapes 5-methylcytosine deamination signatures in pan-cancer genomes” by Bortolini Silveira et al represents an interesting and thorough exploration into several rare mutation signatures thought to be related to 5mC deamination. It uses a variety of experimental systems to provide an in-depth analysis of some of these rare signatures. Of note is the very nice linkage of SBS96 to CpG and to a lesser extent CpA methylation in Uveal Melanoma, as well as linking SBS105 to the rare POLD1 mutation R817W. Whilst this paper will be well suited for publication in Nature Communications, there remains some details which I feel require clarification before acceptance. Overall it was a well-designed and clear study, the results which are of interest to the wider cancer genomics community.

Major comments:
How were signatures defined? How many allowed to be found per tumour? Do I understand correctly that a predefined set of signatures was used as per supp table 4. How was this defined?

How do you know the SBS95 SBS96 and SBS105 are not a misclassified SBS1 signature? Isnt it surprising that most of the tumours with these signatures have 0 mutations from SBS1 whereas every other sample has SBS1 mutations and this is common across all cancer types? These tumours don’t age?

How frequent are these rare signatures compared to other signatures within all mutations in these 12 cancers.
e.g for the SBS105 tumours which have similar total numbers of mutations assigned to SBS105 its 72% and 32% of the total burden. Surely this is relevant. And why is this?

Although Fig 1b addresses this somewhat, the paper is lacking a barplot showing contribution % of these sigs? In all this paper is about 18 patients, and the % mutation burden of these 18 patients should be thoroughly described.

GEL other
SBS95 3/11825
SBS96 7/11825 6
SBS105 2/11825

In supp table 6 there is a list of supposed variant counts but they all have decimal points? I do not understand how they are not whole numbers. What causes this?
38890,42
35678,37

Would be nice to see if you made a POLD1 R817W mutation in a cell line and made a clonal analysis as for the knockout cell lines, would you see the same SBS105 signature after introducing this mutation and clonally expanding a cell? It is also interesting that you state "Noteworthy, this somatic variant corresponds to a CpG>TpG mutation, which we cannot exclude is a consequence of SBS105 mutagenesis" – a chicken and egg problem. Surely it cannot both be responsible for SBS105 and be caused by it? What else do you suggest as the cause of SBS105 mutations? Is there any insight one can gain from looking at the position of this mutation in the structure?

I feel like SBS95 and the novel signature remain unexplained… When you started the story saying "we next sought to explain the mutations not dependent on methylation" and then you explain SBS105 nicely it would be nice to have more of a section about SBS95 and SBS Novel.

Why does the CpA to TpA mutations get classified as SBS96? Shouldn't they be their own signature?

When looking at the CpA methylation in uveal melanocytes what was the extended sequence context? Does it match the DMNT3A motif?

For the cut and run, why does the N terminal construct have 10 fold more peaks?

How do you explain that MDB4 deficiency is responsible for 5mC mutations when MBD4 doesn't seem to bind to methylated DNA? What is the mechanism here? It is clear that knockout of MBD4 increases mutations as well, but if the cut and run doesn't show binding it is a bit strange. A comment about this would be nice.

Minor comments:
Clarify patients, in the methods its says 25 patients but in supp table there are only 6 patients not in GEL. Why weren't the others included in the WGS cohort? When you include different patients for different experiments its hard to understand. It would be nice if there was all experiment types for less patients rather than different patients for different experiments, particularly when you are looking at rare phenomenon.

Make data available, it says its at EGA but what is the accession number?

Fig 1 legend should be 5mC

Fig 3C. Is it correct to adjust the scales between what should be similar results. E.g the N term and C term overexpression constructs a shown to be similar but the scale is adjusted to show this. What is your reasoning for adjusting the scales so they seem similar?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Dear Authors
I honestly and greatly appreciate the extensive work you undertook in a rather short time to answer and elaborate on the reviewers and especially my many, and often admittedly tough, interrogation. I want to apologize for my oversight concerning part of the WGBS methodology and thank for the individual explanations delivered in the rebuttal.
The authors addressed all of my concerns and the added results and controls greatly increase the confidence in the data to support the conclusions. Furthermore, do the presented results not "just" shed light on the clinical observation, but deliver thorough data pointing to new directions and hypotheses that will surely inspire future research about deamination/MBD4-independent C>T transitions and the dynamic (epi)genome.

A few minor points:

The authors write that the MBD4-FLAG proteins are expressed at a comparable level as the endogenous MBD4. The western blot in Fig. S9, clearly shows higher levels of the FLAG-MBD4, which corresponds to about double the level of endogenous protein expression when roughly measured by densitometry and normalized to the ponceau signal. It is a little deceiving – densitometry could be added to the WB. Nonetheless, I don't question the later obtained results.
The authors present a WB analysis for MLH1 (Fig. 5a, 6a), which is not mentioned in the text. I understand this is to control for potential confounding effects of MMR, it should probably be very briefly mentioned.
The authors repeated CUT&RUN assays to control for potential artefact observations, amongst others, due to antibody unspecificity. Unfortunately, this was indeed the case but this observation is critical and commendable. I encourage the authors to maybe include a statement about the (current) infeasibility of CUT&RUN with the used antibodies (e.g. in the

methods).

The authors state that “Notably, while tagged MBD4 enrichment was negatively correlated with CpG methylation percentage” (Fig S10a).
i) could you maybe overlay the plot with the MBD4-signal on a second y-axis, to give a better grasp on how MBD4 binding is distributed over the inspected windows (i.e. is it gradual or not?/ does correspond “lowest” to no enrichment..)
ii) This is rather surprising, given its role as mCpG binding protein and the contrast to the currently accepted view that “highly” or rather more stably methylated CpGs are more prone to spontaneous deamination. Including the above-mentioned data might relativize this apparent discrepancy. If not, I would be looking forward to a short comment about that.

Sincerely, Simon Schwarz, DBM University of Basel

Reviewer #2

(Remarks to the Author)
The authors have satisfactorily addressed all the concerns I had in my prior review. I especially commend them for their exhaustive description of their research methodologies in their revised manuscript.

Reviewer #3

(Remarks to the Author)
This revised manuscript is much improved. The authors have responded to the identified flaws quite well. I am pleased the flaws in the CUT&RUN were corrected and the decision to remove endogenous results also makes sense. I hope to see this amended and published in a later manuscript. I also really liked that the extended context of the CpA mutations matched the DNMT3a motif nicely.

I am happy to accept this manuscript after the resolution of the below comments:

Figures need to define all statistics, what do the box and whiskers of the boxplots represent (median 25th 75th percentile?) what is the shaded region in scatterplots eg 1e

It was commendable that you attempted to recreate the R817W mutation in POLD1 but didn't see any effect. Do you think this could be because there remains sufficient levels of the endogenous wildtype enzyme?

Code availability: Great to see the data deposited, but the analysis involved a lot of code? Shouldn't this also be included to enable recreation of the results? Also source data.

Fig 5d Melanocyte SBS1 is blank

Rebuttal letter

The authors greatly thank the reviewers for their detailed and careful revision of our manuscript, and their contributive comments. We believe that the changes in bioinformatic analyses and presentation of the data prompted by this review considerably improved the overall quality of our manuscript.

In this revised version, the reviewers may notice small changes in values relying on genome-wide CpG methylation calls. Previous analyses were performed without the complete filtering criteria described in the Methods section. This has been corrected throughout the manuscript. The interpretations of the results remain unchanged. Changes in the manuscript's text have been marked in blue color.

Below, we provide point-by-point responses to reviewers' comments, which are also marked in blue color.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors around Bortolini Silveira investigate cancer signatures, focussing on mutations arising from spontaneous Cytosine deamination, which leads to C>T transitions if not repaired. The two DNA glycosylases MBD4 and TDG are of special interest as they are highly efficient in detecting and processing G:T mismatches. While an association of cancer and mutated TDG was only reported in very very few cases, MBD4 defects in cancers are a well-known phenomenon and also previously investigated and reported by the authors.
In the first half (Figures 1 and 2), the authors perform bioinformatical analyses on a subset of single-base substitution patterns (SBS) showing a large contribution of C>T transition, which are derived from cancer whole-genome sequencing data of own and "public" origin. Many correlations are reported based on which the authors support (SBS1, SBS96) or disfavour (SBS95, SBS105, SBSnovel) an association of the C>T transition containing SBS and MBD4 defects. A brief but interesting excursion into CpA methylation and corresponding CpA>TpA mutations in UVM is undertaken, reporting a non-negligible amount such mutations in dependence of MBD4.
In Figure 3, the authors perform experiments in non-related (HAP1) cells to examine the genomic distribution of MBD4 and its relation with several epigenetic characteristics, such as chromatin accessibility and histone methylation. Contrary to the author's expectations MBD4 is abundant in active chromatin regions such as "unmethylated" promoters. Switching back to cancer cells and public annotation-based sites of chromatin activity, they report that protection against C>T transitions is more efficient in transcriptionally active regions. A protective effect in early-replicating areas is also reported.
Finally the authors perform a long-term cultivation experiment to assess the impact of MBD4 and or TDG depletion in HAP1 cells. They found that MBD4 is the dominant enzyme in

protecting against C>T transitions, with a small contribution of TDG, but only in the absence of MBD4. The replication-dependent bias seems to be connected to MBD4.
All in all, the authors provide data that corroborates previous observations and bring some new aspects into the picture.

Major comments:
MBD4 has already been reported several times to be a major contributor in the protection against C>T substitutions and validated in different cancers (especially in SBS1 and SBS96). The novelty factor of this statement is therefore rather low.

Although we agree that *MBD4* deficiency has been extensively linked to an increased CpG>TpG mutational burden in different cancers (including by our team), a comprehensive view of the relationship between *MBD4* deficiency and the different CpG mutational signatures was lacking. By analyzing a previous version of the GEL series, Degasperi *et al*., 2022 (10.1126/science.abl9283) described for the first time SBS96, revealing its contribution in multiple *MBD4*-deficient (*MBD4*def) tumors, but also in a small number of *MBD4* wild-type (*MBD4*wt) tumors. Here, we comprehensively analyzed the contribution and statistical significance of SBS96 signature fitting in a new release of the GEL series, in addition to multiple WGS datasets on *MBD4*def tumors of tissue origins not included in GEL. This allowed us to show that high statistical confidence SBS96 was exclusively found in *MBD4*def tumors, and that all *MBD4*def tumors showed predominant SBS96 contribution (Fig. 1b,c; Supplementary Table 8). Meanwhile, high CpG>TpG mutational burden in *MBD4*wt tumors could be well explained by common SBS1 or by the potential new rare signature SBSnovel. Hence, we describe for the first time that SBS96 is intrinsically associated with *MBD4* deficiency in tumors of multiple tissue origins. We now more clearly state our findings in the Results section (lines 107-121).

Finding two (to three) more signatures that could be de novo associated to MBD4 is rather interesting, the authors however reject this possibility based on rather superficial and correlative analyses. For example, a (not so big) strand-bias in SBS95 and SBSnovel, compared to no strand-bias in SBS1 and SBS96, is not enough ground to discard the possibility of MBD4-dependence, especially since no (mechanistic) analyses concerning genomic region, transcriptional activity, or actually replication timing etc. are performed.

We thank the reviewer for warning us about the suboptimal data analyses to fully support our claims. Therefore, we now include additional analyses showing that transcriptional strand asymmetries of SBS95, SBSnovel, and SBS105 are largely dependent on gene expression levels, being strongest in highly expressed genes and absent in lowly expressed genes (Fig. 1f; Supplementary Fig. 3c). To our knowledge, 5mC deamination repair has not been described as coupled to the transcription machinery, and the absence of transcription strand asymmetry in signatures SBS1 and SBS96 is in agreement with current literature (Fig. 1f). However, strand asymmetry in SBS95 and SBSnovel is indicative of the role of TC-NER instead. This is consistent with their lower mutation rates in highly expressed genes (Fig. 1f), where TC-NER is expected to be more prominent. Although we cannot completely discard some contribution of 5mC deamination in SBS95 and SBSnovel, our data indicates that these signatures cannot be

fully explained by 5mC deamination alone. Notably, SBS95 and SBSnovel higher C>T mutation rates in the transcribed strand rather suggest a preferential accumulation of G>A substitutions in the untranscribed strand. It is therefore possible that certain types of DNA damage directly targeting guanines are involved in these signatures. We rephrased the Results (lines 160-176) and Conclusions (lines 447-452) to more clearly state the interpretation of our findings.

While, I agree that the pattern of SBS105 does not (solely) link to 5mC deamination, as it is rather clear given the additional enrichment of other than C>T transitions (Fig 1a). The non-linear but quite distinct correlation between mCpG levels and CpG>NpG mutations, could still point to a mC-dependent mechanism, potentially involving dynamic mC oxidation by TETs and finally TDG-initiated BER. A clear exclusion of mC-dependence, as also mentioned in the discussion, is not appropriate. Based on this, the authors should either provide more rigorous analyses/experiments or need to adapt their rather definitive phrasing ("SBS1 (age) and SBS96 (MBD4 deficiency) are the sole 5mC deamination signatures") accordingly.

We greatly appreciate the reviewer's comments, as they prompted us to restrict Fig. 1d-f to CpG>TpG mutations, allowing now a better comparison between the different signatures. While we agree that the partial association of SBS105 with the overall genomic distribution of CpG methylation could point to a 5mC-dependent mechanism, CpG>TpG mutated sites in SBS105 are rather found hypomethylated at the single cytosine level. This is in sheer contrast with the other CpG signatures analyzed (Fig. 1d). We now present a comprehensive view of the different CpG substitution classes of SBS105 in Supplementary Fig. 3a,b, showing that the poor association of SBS105 with CpG methylation is equally observed from C>A, C>G, and C>T substitutions. The presence of transcription strand asymmetry for all major substitution classes of SBS105 (Supplementary Fig. 3c) further indicates that 5mC deamination alone cannot fully explain this signature. Finally, we now propose that SBS105 CpG>NpG mutagenesis might arise from nucleotide misincorporation at least partially following TC-NER. Following the reviewer's suggestion, we rephrased the Results (lines 177-184) and Conclusions (lines 443-447) sections to better reflect our findings, more precisely stating that 5mC deamination is unlikely to play a major role in SBS105 and that other mutagenic processes are most probably involved.

In general, the consistency of analyses performed and conclusions drawn is sometimes questionable. The manuscript would gain a lot of trust as well as strength if more analyses would be performed around the observed patterns and phrasing would be adjusted. E.g.

- While the effect of increased CpA>TpA mutations in UVM cells is quite distinct and interesting, the authors should provide more structured comparisons. mCpA levels (Fig 2f) should, if available, also be compared between cancer tissues (analogous to the comparison in Fig2b), if possible stratified per cancer type, especially for the HGG/Oligodendrocyte-like tissue harbouring higher levels of mCpA.

We thank the reviewer for the comments above, as they led us to include additional whole genome bisulfite sequencing (WGBS) on three metastatic UVM samples. Despite our best efforts, we could not access WGBS data on HGG in a timely fashion to be included in our manuscript. Our new WGBS dataset is nevertheless quite informative and supports our claims. Firstly, we confirm that CpA methylation is detectable in UVM cells, with the highest non-CpG methylation levels in CAC contexts in the three metastatic UVM samples. Secondly, one of the UVM samples showed CAC methylation at a similar level to normal uveal melanocytes (Supplementary Fig. 6), indicating that transformed UVM cells can maintain high levels of CpA methylation. The lower but still distinct CAC methylation level in the remaining two UVM samples may result from higher proliferation rates in these tumors, as both samples showed some evidence of methylation degradation in CpGs prone to hypomethylation due to cell division (Zhou *et al*., 2018, 10.1038/s41588-018-0073-4) (Supplementary Fig. 6). Our results are thus consistent with CpA methylation being generally inversely correlated with the rate of cell division. Overall, we show that UVM cells can sustain surprisingly high levels of CpA methylation, a feature not expected to be prevalent in highly proliferative cancer types.

-Although high methylation levels don't necessarily mean stable mC levels (https://doi.org/10.1038/s41467-020-16354-x). The hypothesis of more stable, and therefore more prone to deamination, mCpA in OVM cells due to lower turn-over by TETs (and TDG) is intriguing. It would however be more telling if the absolute expression values, as depicted in Fig 2g, would i) be separated into MBD4def and MBD4wt UVMs and ii) also be put into relation with TET expression levels in the corresponding healthy tissue to asses this potential causation. Also, this should be related to the observation that C>T transitions generally happen at highly methylated, i.e. potentially more stably methylated, sites (Supp Fig 10).

We thank the reviewer for the suggested reference, which is now cited in the Results section. Although we agree that the separate analysis of expression values by *MBD4* status would be more telling, *MBD4*def tumors are rare and the TCGA series only includes two such tumors. Hence, comparing expression levels between *MBD4*wt versus *MBD4*def tumors of various tissue origins would not be presently contributive.

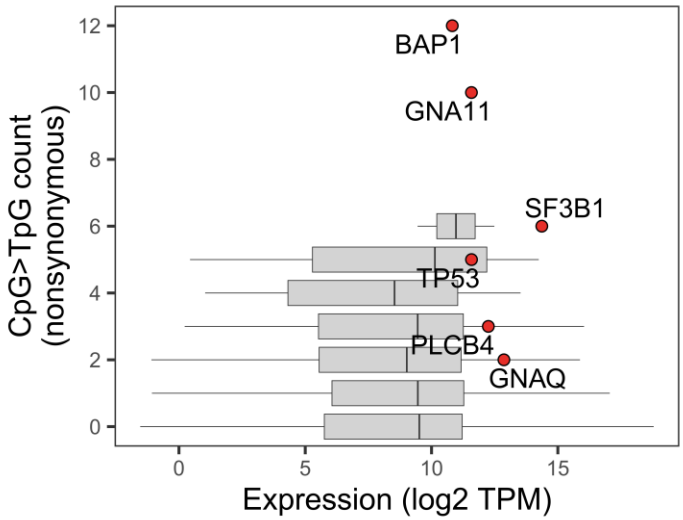
Following the reviewer's suggestion, we now include the analysis of gene expression differences among most of the normal cell types mentioned in the manuscript (Fig. 3i). This analysis is derived from single-nuclei RNAseq data of the ocular posterior segment (Monavarfeshani *et al*., 2023, 10.1073/pnas.2306153120), which includes expression data on uveal melanocytes, oligodendrocytes, and less differentiated oligodendrocyte precursor cells (OPCs). Of note, most HGGs are believed to arise from less differentiated glial progenitors instead of fully differentiated post-mitotic glial cells (Alcantara Llaguno *et al*., 2016, 10.1038/bjc.2016.354). Firstly, we show that OPCs express lower levels of *DNMT3A* than fully differentiated oligodendrocytes without major changes in the expression of *TET* genes, indirectly indicating that *de novo* CpA methylation is modulated throughout glial differentiation. Secondly, we show that uveal melanocytes express higher levels of *DNMT3A* and lower levels of *TET* genes than OPCs, indirectly indicating that uveal melanocytes might accumulate higher levels of CpA methylation than certain glial precursors. Hence, the sole fact that fully differentiated

oligodendrocytes show higher CpA methylation levels in comparison to uveal melanocytes (Fig. 3h) is not contradictory with the higher CpA>TpA mutational burden in *MBD4*def UVM versus MBD4def HGG.

We also included in the Results section a direct comparison between CpGs (more stably methylated at high percentages due to the activity of DNMT1, but with fewer sites in the genome) versus CpAs (less stably methylated, but much more abundant in the genome). We show that the absolute number of methylated CpGs in uveal melanocytes was ~9 fold higher than that of methylated CpAs, which is consistent with a CpG>TpG mutation rate ~9.5 fold higher than that of CpA>TpA in *MBD4*def UVM. Hence, regardless of the variable dynamic rates of DNA methylation and demethylation in different sequence contexts, we believe it is the overall abundance of methylated sites in each context over time is the main factor dictating the relative frequencies of different C>T mutations found in *MBD4*def tumors.

- To gain deeper information about “how”, rather than just “that” MBD4 is such a frequent cancer driver, I suggest to translate expression analyses to the data in Figure 2h, where i) transcriptional levels of these genes could be correlated to the mutation rate (are these genes generally highly expressed in the corresponding UVM types) and ii), more importantly, whether these sites, where mutations are observed, show a high mC level or not, akin to what was performed by Robertson et al. (REF55), although not regarding the state of MBD4 or CpA. Potential differences or similarities between effects at CpA and CpG sites could be highlighted.

We greatly thank the reviewer for these suggestions, which prompted us to further explore the mechanisms underlying the targeting of oncogenic drivers in *MBD4*def UVM tumors. We now show that genes more frequently targeted by nonsynonymous CpG>TpG mutations tended to show a higher number of methylated CpGs in the CDS (Fig. 4c). Gene expression by itself had little effect on recurrency (shown below).



Most importantly, our data shows that frequent targeting of *BAP1*, *GNA11*, *SF3B1* and *TP53* could not be explained by either their abundance of mCpGs in the CDS or by their expression levels (Fig. 4c,d), suggesting a strong selective pressure for these mutations. The preferential

accumulation of hotspot mutations associated with UVM transformation in *BAP1*, *SF3B1*, *GNAQ*, *GNA11,* and *PLCB4* (Fig. 4b) reinforces this interpretation. We also show that CpG sites harboring oncogenic mutations specifically in UVM, AML or other tumor types are found largely methylated in any of the normal cell lineages analyzed (Fig 4e), indicating that targeting of tumor type-specific genes is most probably caused by tissue-specific positive selective pressures and not by tissue-specific DNA methylation patterns.

Although we agree that differences or similarities between effects at CpA and CpG sites could be highlighted, the small number of CpA>TpA mutations in genic regions prevented us from exploring CpA mutations with coding consequences in greater detail. Nevertheless, two oncogenic CpA>TpA mutations were found in *BAP1* in *MBD4*def UVM, as shown in Fig. 4a.

- Comparisons between the replication timing and examined chromatin marks (Fig. 3c) should be made, to estimate the connection of replication timing, accessibility, methylation state and finally mutation rate.

We thank the reviewer for this suggestion. Mutation rates were calculated based on chromatin states annotations, which are defined by the combination of the 6 histone marks now presented in Fig. 5b. We now show in Supplementary Fig. 11 the overlap between chromatin states in each lineage with the pan-tissue replication timing annotations. As expected, early replicating regions are found highly enriched in active chromatin states, whereas late replicating regions show enrichment of low/quiescent and heterochromatin states. This was similarly observed for the five cell lineages analyzed. Based on this analysis, we noticed that the previous breast epithelial lineage chromatin states annotation from ENCODE had an overrepresentation of low/quiescent states (regions without any histone marks), indicating poor annotation quality. Therefore, a higher-quality ENCODE annotation from the same breast epithelial lineage was used instead in the revised version of our manuscript, which increased consistency with the other lineages analyses. Importantly, our conclusions remain unchanged.

- It is not entirely surprising that MBD4 is binding not only to methylated promoters, as first, it is able to bind carboxylcytosine (10.1093/nar/gks714), which is not detected in a bisulfite sequencing, and second, its glycosylase activity is pretty much independent from the methylation state in the vicinity (https://doi.org/10.1128/MCB.00588-08). Even more so, it was already published that MBD4 frequently binds to promoters that, not only show mC abundance but several marks of active transcription (https://doi.org/10.1016/j.cell.2013.03.011). Furthermore, the analysis the authors use does not provide the resolution (with figure 4c or Supp 8) to deny that MBD4 is actually binding to localized mCpGs in an otherwise low-methylated promoter. Given the previous observations, it would be interesting to see the mCpA levels in HAP1 cells and a corresponding graph of mCpA levels in Fig 3c.

Unfortunately, the documentation of bioinformatical analyses and datasets generated, as well as information/visualization of quality control is insufficient (public datasets exempt):

- TDG and MBD4 CUT&Run datasets are not findable/accessible, and major information about the cell lines generated is missing. The “validation” immunofluorescence in Supp Fig. 11 has to be mentioned earlier, and obvious unspecificities cannot be neglected. E.g. uncropped western of potential degradation fragments in N-terminally tagged MBD4 could explain such a high number of peaks reported in Fig 3a. No information about number of replicates, sequencing depth, peak calling parameters (e.g. FDR and fold enrichment) are mentioned. Coverage tracks with higher magnification than Supp Fig 8. should be provided and/or a signal vs. Noise estimation on in the TDG experiment, as all available antibodies struggle with unspecificities.

We greatly appreciate the reviewers’ expertise and appropriate questioning of our CUT&RUN data, which has led us to make major adjustments in both our experimental and bioinformatics approaches. We describe these changes in detail below:

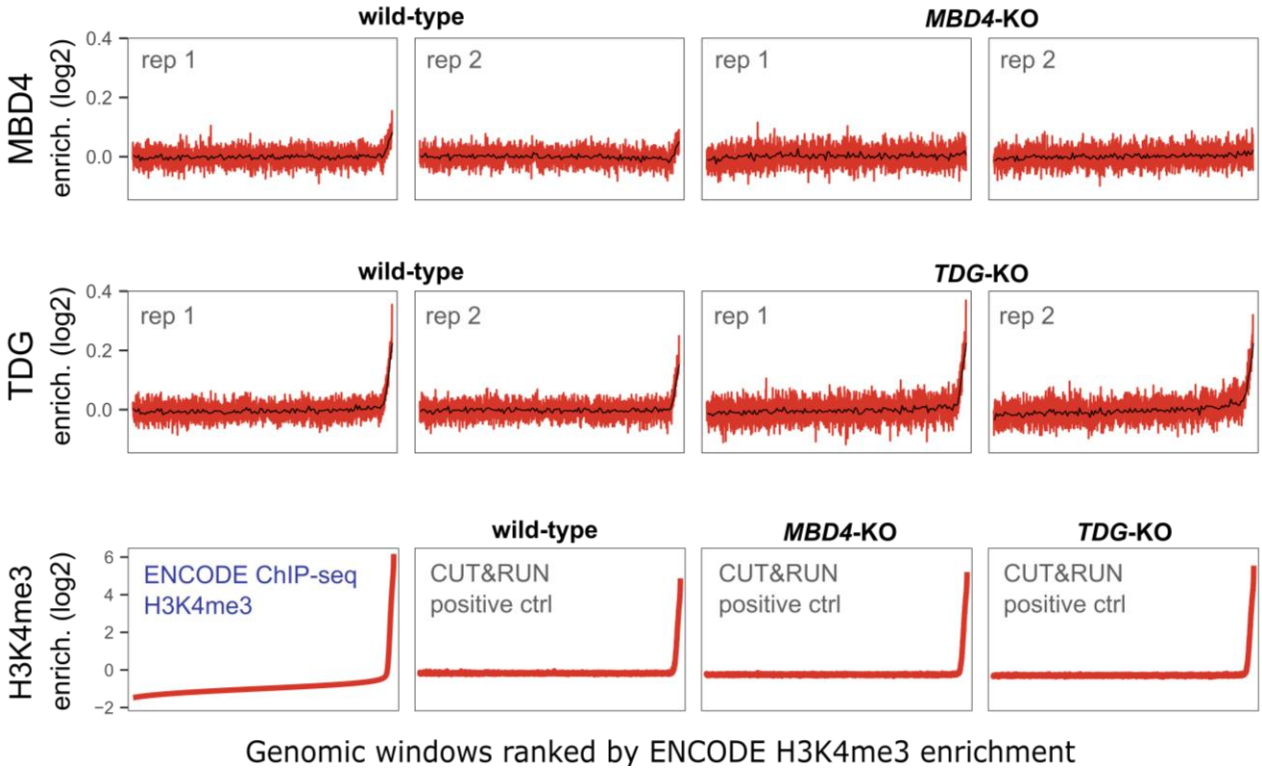
- All CUT&RUN experiments were repeated with an optimized version of the EpiCypher kit (CUTANA ChIC/CUT&RUN Kit V4). With this modification, we observed a better consistency of our replicates due to increased chromatin recovery, and a lower background in open chromatin regions due to a better selection of mononucleosome-sized chromatin fragments before library prep.

- CUT&RUN on tagged MBD4 was originally performed on bulk cell populations expressing highly variable (Supplementary Fig. 9a) and often non-physiological levels of MBD4. We thank the reviewer for bringing to our attention the possibility of degradation fragments in N-terminally tagged MBD4, which was indeed the case for clones with high exogenous expression. We thus selected single-cell clones expressing N- or C-terminally tagged MBD4 at levels comparable to endogenous MBD4, as confirmed by western blotting (Fig. 5a). No major degradation products of as low as 10 kDa were observed for either clone (Supplementary Fig. 9b). We then obtained FLAG enrichment by CUT&RUN in each tagged MBD4 clone in duplicates, and in parental HAP1 cells as a negative control. Overall, we believe this new experimental design improved our ability to confidently map genome-wide enrichment of tagged MBD4 expressed at biologically relevant levels.

- Baubec *et al.*, 2013 (10.1016/j.cell.2013.03.011), cited by the reviewer, observed that conventional peak calling was largely incompatible with their ChIP-seq data on tagged MBD4 expressed in mouse embryonic stem cells. We indeed observed the same limitation with our newly generated CUT&RUN data. We thus followed Baubec et al’s analytic strategy based on genomic windows. We binned the genome into 2-kilobase windows and selected the highest- and lowest-ranked windows based on the enrichment of different histone marks ChIP-seq from ENCODE (Fig. 5b). Tagged MBD4 signal enrichment was positively correlated with activating histone marks H3K4me3 (promoters), H3K27ac (active promoters and enhancers) and H3K4me1 (active and primed enhancers), and to a lesser extent to H3K36me3 (transcribed gene bodies). Conversely, tagged MBD4 signal enrichment was negatively correlated with repressive marks H3K27me3 (polycomb repressed) and H3K9me3 (heterochromatin) (Fig. 5b,c). Overall, we confirm our previous observation that both C- and N-terminally tagged MBD4 preferentially bind to active chromatin domains. In addition, we now show that while tagged MBD4 enrichment is negatively correlated with CpG methylation percentage, it is positively

correlated with both CpG and mCpG densities (Supplementary Fig. 10). Hence, our data supports the preferential binding of MBD4 to mCpG-rich genomic regions.

- CUT&RUN on endogenous MBD4 and TDG were repeated using HAP1 cells wild-type and knockout for each gene, in duplicates. We observed higher endogenous MBD4 signal enrichment in H3K4me3 highest-ranked windows (shown below), in agreement with our tagged MBD4 data. This pattern was apparent in both replicates performed in wild-type cells, but not in replicates performed in *MBD4* knockout cells. However, this experiment suffered from poor sensitivity, and we believe it would add little value to our revised manuscript. The experiment on endogenous TDG showed poor specificity instead, with similarly high TDG enrichment in H3K4me3 highest-ranked windows in wild-type and *TDG* knockout cells (shown below). Notably, H3K4me3 positive control reactions performed in cells of the three genotypes showed equally strong signal enrichment in ENCODE ChIP-seq H3K4me3 highest-ranked windows. We believe that the limited sensitivity and/or specificity of these experiments were largely dependent on the quality of the antibodies currently available. Hence, we opted to withdraw CUT&RUN data on endogenous MBD4 and TDG from our manuscript. Despite this technical limitation, we believe that the overall conclusions of our study remain largely unchanged. Firstly, we show that both N- and C-terminally tagged MBD4 preferentially bind to active chromatin domains, in agreement with the literature (Baubec *et al.*, 2013, 10.1016/j.cell.2013.03.011). Secondly, it has been previously shown that tagged TDG preferentially binds to active chromatin domains rich in H3K4me3 in mouse embryonic stem cells (Neri *et al.*, 2015, 10.1016/j.celrep.2015.01.008).



- Regarding the analysis of CpA methylation in HAP1 cells, it is important to point out that these cells have a very short doubling time (~14h) and are not expected to accumulate significant levels of mCpA. Hence, we believe this is not a relevant model for this type of analysis.

- Newly generated raw and analyzed CUT&RUN data are now deposited in GEO under the accession GSE275181. This information has been included in the manuscript. The Methods section now includes replicate and sequencing depth information, in addition to details on data analysis based on genomic windows (lines 725-727, 733-744). Mapping statistics are now shown in Supplementary Table 14.

- Apart from the cropped western blot in Fig 4a, no information about the newly established cell lines are given. This should at least include a display of verification PCRs on DNA and cDNA as well as duplication times of these clones (optimally monitored over the 120 day window of cultivation).

As suggested, we now present an extended characterization of knockout cell lines. Genomic DNA PCR confirmed the deletion of the region spanning *TDG* exons 2-5 (Supplementary Fig. 13a), leading to a significant decrease in *TDG* mRNA expression, as measured by qRT-PCR with primers spanning exons 7-8 (Supplementary Fig. 13b). Uncropped western blotting images illustrate the absence of truncated forms of MBD4 (with antibodies against C- or N-terminus) or TDG in knockout cells (Supplementary Fig. 13c). As suggested, we measured doubling times of each cell clone at D60 of long-term culturing, in biological triplicates. Doubling time differences were minimal and non-significant (Supplementary Fig. 13d).

- No detailed information about the WGS sequencing is given for these cells. How many replicates per clone? How many reads, how high was the coverage? Some of those details are given for the patient-derived samples but it is not visible whether this is applicable to the HAP1 cells.

All this information is in the Methods' section entitled "HAP1 long-term culturing and Whole-Genome Sequencing". This section includes details on number of subclones analyzed per genotype, DNA extraction, library preparation, and sequencing coverage. For more clarity, we have changed the titles of the Methods sections that detail the sequencing of patient samples.

- The authors explanation on how the replication timing data was employed is very limited. I assume their mentioning of "pan-tissue" replication timing refers to regions which display similar RT among different tissue types. The authors need to elaborate on this and discuss how many sites they examined, and whether the examined tissues are part of that collection.

The replication timing genomic annotations used were derived from a wide range of cell types and differentiation intermediates of human development (Poulet *et al*., 2019, 10.1093/bioinformatics/bty957). This annotation covers ~85% of the human genome, which is subdivided into constitutive early, constitutive late, or dynamic replication timing during cell differentiation. These descriptions are now provided in the Results and Methods sections. Most

importantly, a pan-tissue annotation was effective in revealing lower relative mutation rates in early replicating regions regardless of tumor type analyzed (Supplementary Fig. 12).

- In the analysis of DNA methylation blocks, the authors should provide the following information to get a better understanding of the data: the number of blocks retained for the analysis, the number of promotors /regulatory elements covered by these blocks and the (average) amount of CpGs in the blocks covered.

As suggested, we included in the Results section more detailed information on CpG blocks. We selected for analysis 698,467 high-quality methylation blocks overlapping promoters (27,698 distinct genes), exonic regions (28,148 distinct genes), enhancers and DNase hypersensitive sites (402,977 cis-regulatory elements from ENCODE). On average, retained blocks spanned a length of ~558 bp and ~11 CpGs.

Minor comments:
In Fig 1a, the authors should label the x-axis with an exemplary “sequence context” and might highlight the most frequently enriched C>T transitions, as done in Fig 2b.

The figure was adapted accordingly.

Although written in the text, it would be handy to have a (table-like) overview of how many samples/patients are represented in the individual SBS categories, and how many of those carry which (known) mutation. Otherwise it is difficult to estimate the extent of observed results.

We appreciate this suggestion. We adapted Fig.1a accordingly to include the number of cases harboring each rare mutational signature and each known mutation. A similar table-like overview describing mutagenic processes and DNA repair pathways potentially involved in each signature was also included in Fig. 1f.

Fig2. While the brain has provided multiple and early evidence for CpA methylation, the references for non-CpG DNA methylation are a little dated and, therefore the statement: “CpA methylation primarily accumulates in differentiated tissues of neural origin by DNMT3A-mediated de novo methylation42,43,44” should be adapted to incorporate more (recent) observations about non-CpG methylation, e.g. in stem cells, spleen, etc.

We acknowledge our previous writing was misleading. This has been corrected and we now reference a more recent review focused on non-CpG methylation (Jang *et al*., 2017, 10.3390/genes8060148).

The authors mention that they “observed a similarly strong TDG binding in active TSSs and active enhancers (Fig. 3c; Supplementary Fig. 8)”, which is not an appropriate statement given the mentioned two figures. While an enrichment of TDG and MBD4 at TSS is

visible, the signal(intensities) are not quantifiable/comparable through heatmaps, and the binding strength can most certainly not be assessed by ChIP/CUT&Run assays. A comparison of binding frequency can only be estimated by densitiy plots (integrated from the heatmaps), but actual quantitative comparison between MBD4 and TDG binding would require multiple level of normalization, including Antibody and chromatin spike-in methods, aside from other biochemical experiments for binding strength.

We fully agree with the reviewer's worries regarding our previous phrasing. Considering we opted to withdraw our CUT&RUN data on endogenous MBD4 and TDG, this statement has been removed from the Results section.

Fig3e: The authors mention a relative mutation rate. Relative to what? Also, the legend hints to two different significance values (**<0.01 and ****<0.0001) while in the graph both comparisons show ****.

We thank the reviewer for identifying the unprecise description of the relative mutation rates now presented in Fig. 5e,f. Mutation rates observed relative to expected were used, considering an expected random distribution of CpG>TpG mutations among mCpG of each lineage. This description has been included in the figure legends. The reference to P-value <0.01 was removed, as it was not relevant to this figure.

The authors mention that "[...] was significantly more pronounced in SBS1 MBD4wt cases than in SBS96 MBD4def cases (ratio of repressed mutation rate by active mutation rate: 3.77±1.20 [SBS1] vs 1.56±0.30 [SBS96]; Wilcoxon t-test P = 1.27e-11)".
They compare two different signatures in two different genotypes. I believe that the same signatures should be compared over the genotypes, and the data should be visualized.

We thank the reviewer for the suggested analysis, but we believe it is technically unrealistic because: (i) SBS96 is exclusively found in *MBD4*def tumors and a direct consequence of this DNA repair deficiency; (ii) SBS1 was only found in 3 *MBD4*def tumors at a minor percentage contribution (Supplementary Fig. 1). In practice, each CpG>TpG mutation in these tumors has a higher probability of being derived from SBS96 than from SBS1. Hence, it is challenging to isolate SBS1 mutations in *MBD4*def tumors.

While sometimes indicated in the text, statistical indicators are missing for the whole Fig 4.

We recognize the previous lack of statistical indicators comparing each genotype separately, which was a direct limitation of having analyzed only two subclones per genotype. To address this, we performed WGS on additional subclones, for a total of four subclones analyzed per genotype. Statistical indicators are now included in Fig. 6c,e.

Reference 27 is corrupted

This reference was corrected.

Reviewer #2 (Remarks to the Author):

In this study, the authors seek to validate the etiology of rare COSMIC mutation signatures using a wide cohort of tumor data sets spanning over 12000 genomes, specifically focusing on the 5-methylcytosine deamination-based signatures. 5-mC deamination in CG contexts generates C>T transitions, which are among the most abundant mutations in the mutational spectra of a majority of tumors; however, the relationship between 5-mc deamination and C>T centric SBS signatures is poorly understood. The authors observed a striking correlation between the prevalence of C>T changes and the expression of DNA glycosylases, particularly MBD4. A subset of MBD4 deficient tumors had a unique remarkable enrichment of the rare SBS96 signature, which previously had a somewhat confounding etiology. Further, the authors have done an adequate job of combining genome wide methylation and histone modification data with signature activities to understand the distribution of C>T mutations across the genome. Finally, the authors were able to recapitulate the contribution of glycosylases to the observed signatures using HAP-1 cell lines with knockouts for both the major 5mC-deamination glycosylases MBD4 and TDG.

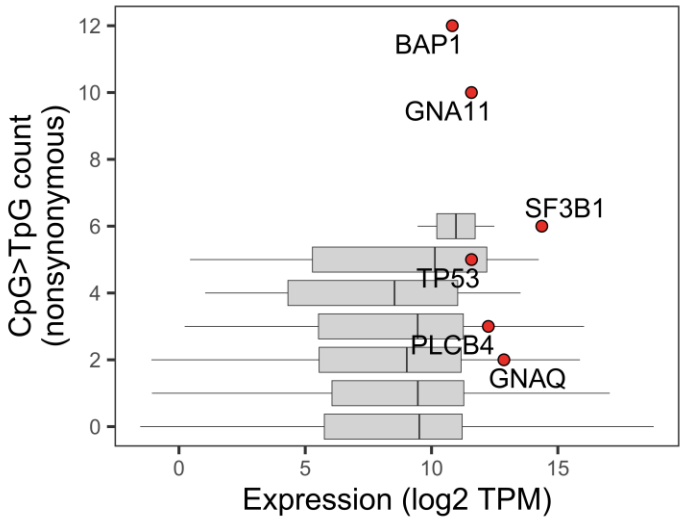
Minor comments:

1) The contribution of MBD4 to CG>TG burden in UVMs is quite interesting. Have the authors considered a correlation between with C>T mutations generated from UV radiation, and see whether or not there is a significant overlap? In line with this, what proportion of the observed C>T mutations in MBD4wt and deficient tumors CC>TT dinucleotide changes

We thank the reviewer for this question, as it brought to light the necessity to better describe the particularities of UVM regarding exposure to ultraviolet radiation. Features linked to ultraviolet radiation (mutational signatures SBS7a/b and DBS1 combined with a high tumor mutation burden) are largely restricted to cutaneous melanoma (Degasperi *et al*., 2022, 10.1126/science.abl9283). In UVM, these features are only observed in the rare tumors of the iris, which is located anteriorly within the uveal tract and is directly exposed to sunlight. In contrast, ultraviolet radiation features are absent in UVM of the choroid and ciliary body (Johansson *et al*., 2020, 10.1038/s41467-020-16276-8). All *MBD4*def UVM analyzed in this manuscript were choroidal and did not show any SBS7a/b contribution. Hence, for the comparison of CpA>TpA mutation rates in UVM, we focused on choroidal tumors without any evidence of ultraviolet radiation damage. We modified the Results and Methods sections accordingly to precisely state this information.

2) It was also perhaps surprising that there was no overlap between the oncogenic driver mutations between any of the MBD4def tumors. While tumor-specific selection can explain this to some extent, I'm curious if there are additional factors. Can you stratify the MBD4wt v MBD4def signatures by age across various tumor types? Perhaps there is some clock-like component to mutations observed in the absence of MBD4.

We greatly thank the reviewer for these suggestions, which prompted us to further explore the mechanisms underlying the targeting of oncogenic drivers in *MBD4*def UVM tumors. We now show that genes more frequently targeted by nonsynonymous CpG>TpG mutations tended to show a higher number of methylated CpGs in the CDS (Fig. 4c). Gene expression by itself had little effect on recurrency (shown below).

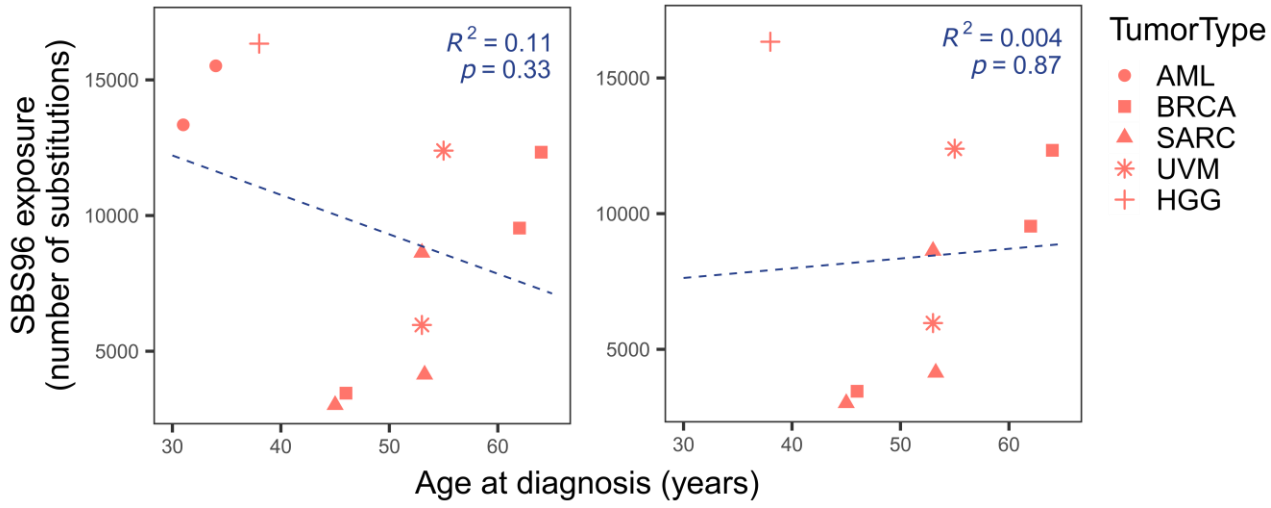


Most importantly, our data shows that frequent targeting of *BAP1*, *GNA11*, *SF3B1* and *TP53* could not be explained by either their abundance of mCpGs in the CDS or by their expression levels (Fig. 4c,d), suggesting a strong selective pressure for these mutations. The preferential accumulation of hotspot mutations associated with UVM transformation in *BAP1*, *SF3B1*, *GNAQ*, *GNA11,* and *PLCB4* (Fig. 4b) reinforces this interpretation. We also show that CpG sites harboring oncogenic mutations specifically in UVM, AML or other tumor types are found largely methylated in any of the normal cell lineages analyzed (Fig 4e), indicating that targeting of tumor type-specific genes is most probably caused by tissue-specific positive selective pressures and not by tissue-specific DNA methylation patterns.

Regarding the overlap in oncogenic driver mutations found among different tumor types, it is important to mention that most of our data on recurrency relied on *MBD4*def UVM, from far the tumor type with the highest prevalence of MBD4 mutations. UVM shows a unique landscape of frequently mutated drivers (Johansson *et al*., 2020, 10.1038/s41467-020-16276-8). As an example, oncogenic mutations in either *GNAQ* or *GNA11* are highly prevalent in UVM but are extremely rare in other tumor types (Larribère *et al*., 2020, 10.3390/cancers12061524). Hence, we believe the limited overlap between oncogenic mutations in UVM versus other tumor types was expected. Notably, we found CpG>TpG mutations in *TP53* in multiple tumor types,

including UVM. Despite our best efforts to gather or generate data on as many *MBD4*def tumors as possible, their rarity limits the feasibility of analyzing recurrency in other tumor types.

Although we agree that both SBS1 and SBS96 may show a clock-like behavior, it is essential to point out that these signatures emerge in different contexts. SBS1 is the result of leaky 5mC deamination repair in otherwise BER-proficient cells, a process that takes place during fetal development and throughout adult life. Hence, it is expected that SBS1 mutational burden correlates with patient age at diagnosis in multiple tumor types (Alexandrov *et al.*, 2015, 10.1038/ng.3441). In contrast, MBD4 deficiency is most often acquired in somatic cells due to the loss of the *MBD4* wild-type allele in individuals harboring heterozygous germline *MBD4* loss-of-function mutations. This event takes place at unknown time points during development, and SBS96 mutational burden might indicate instead the time from *MBD4* biallelic loss until tumor cell clonal expansion. The exceptions are the two AML cases found in siblings harboring germline biallelic loss of *MBD4* (Sanders *et al.*, 2018, 10.1182/blood-2018-05-852566). As shown below (for patients we had access to age at diagnosis), SBS96 exposure was not correlated with patient age at diagnosis regardless of the inclusion or exclusion of the two AML cases.



3) The choice of using diploidized HAP1 cells for clonal expansion experiments isn't adequately explained. Please elaborate in the methods section.

We thank the reviewer for this suggestion. Haploid HAP1 cells tend to diploidize over time, potentially altering relative mutation rates per genomic position at unknown time points during the long-term culturing. To avoid this confounding factor, we opted to analyze diploid HAP1 isogenic cell lines. We included this statement in the Methods section.

Reviewer #3 (Remarks to the Author):

This paper titled “Base-excision repair pathway shapes 5-methylcytosine deamination signatures in pan-cancer genomes” by Bortolini Silveira et al represents an interesting and thorough exploration into several rare mutation signatures thought to be related to 5mC deamination. It uses a variety of experimental systems to provide an in-depth analysis of some of these rare signatures. Of note is the very nice linkage of SBS96 to CpG and to a lesser extent CpA methylation in Uveal Melanoma, as well as linking SBS105 to the rare POLD1 mutation R817W. Whilst this paper will be well suited for publication in Nature Communications, there remains some details which I feel require clarification before acceptance. Overall it was a well-designed and clear study, the results which are of interest to the wider cancer genomics community.

Major comments:
How were signatures defined? How many allowed to be found per tumour? Do I understand correctly that a predefined set of signatures was used as per supp table 4. How was this defined?

Indeed, we used a predefined set of signatures collectively described in the landmark work by Degasperi *et al*., 2022 (10.1126/science.abl9283). The definition of signatures used, including their trinucleotide context frequencies and their categorization as common or rare in different tissue types is available from the R package *signature.tools.lib*, which was made available to the research community by their group. A maximum of a single rare signature was allowed per sample, a default setting of the package’s *FitMS* function. We now explicit this information in the Methods section (lines 595-596). Of note, the categorization of common and rare signatures per organ limits the possibility of overfitting arising from the large number of SBS signatures available. Considering that these definitions may be updated in the future, we opted to provide the precise lists of signatures used in this manuscript as Supplementary Tables.

How do you know the SBS95 SBS96 and SBS105 are not a misclassified SBS1 signature? Isnt it surprising that most of the tumours with these signatures have 0 mutations from SBS1 whereas every other sample has SBS1 mutations and this is common across all cancer types? These tumours don’t age?

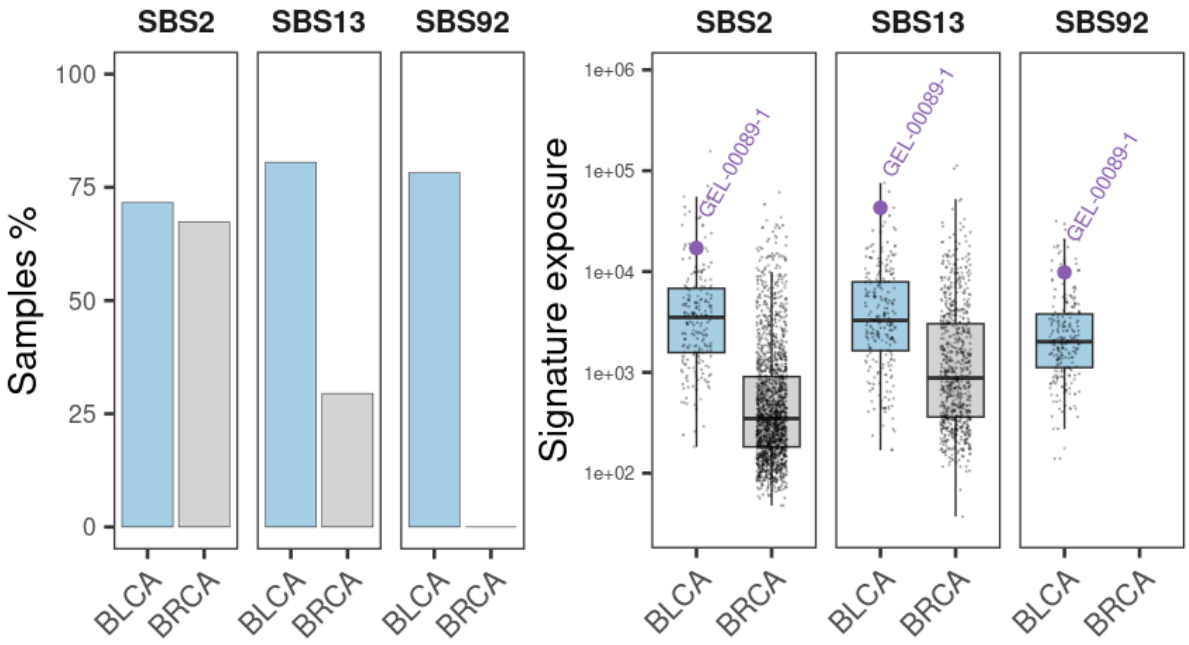
We appreciate the reviewer’s comments, as they reflect some of our hypotheses when we began to analyze the different CpG mutational signatures. Although these signatures are primarily defined by their characteristic substitution frequencies per trinucleotide context (Fig. 1a), our data analysis uncovered additional features that separate them from SBS1: (i) SBS96 is exclusively found with high statistical confidence in *MBD4*-deficient tumors (Fig. 1c, Supplementary Table 8); (ii) SBS95 and SBSnovel show transcription strand bias towards the transcribed strand, which we now show is highly dependent on gene expression levels (Fig. 1f). To our knowledge, 5mC deamination repair has not been described as coupled to the transcription machinery, and the absence of transcription strand asymmetry in signatures SBS1 and SBS96 (Fig. 1f) is in agreement with current literature. Strand asymmetry in SBS95 and

SBSnovel is indicative of the role of TC-NER instead; (iii) SBS105 CpG>NpG mutated sites are found rather hypomethylated in their corresponding cell lineages, which is in sheer contrast with the other CpG signatures analyzed. The presence of transcription strand asymmetry for all major substitution classes of SBS105 (Supplementary Fig. 3c) further indicates that 5mC deamination alone cannot fully explain this signature. Finally, we now propose that SBS105 CpG>NpG mutagenesis might arise from nucleotide misincorporation at least partially following TC-NER. Overall, the presence of unique genome-wide mutational patterns in each rare CpG mutational signature strongly indicates they indeed represent separate entities well distinguishable from SBS1 and with different potential causes than aging.

The absence of SBS1 detection in most tumors harboring rare CpG mutational signatures (Supplementary Fig. 1) most probably relates to the inherent sensitivity limitation of the signature fitting algorithm in hypermutator contexts. A default setting of the *FitMS* signature fitting function is a threshold of 5% minimum signature exposure, which limits the possibility of overfitting. We now explicit this information in the Methods section (lines 595-596). In hypermutated cases with an overwhelming contribution of a single rare mutational signature (e.g., SBS96 and SBS105), it is probable that the common SBS1 signature is dropped out of the final fitting solution due to its low percent contribution. Therefore, SBS1 non-detection does not mean absence of aging in these hypermutated tumors, but rather an unsignificant contribution in percentage.

How frequent are these rare signatures compared to other signatures within all mutations in these 12 cancers.
e.g for the SBS105 tumours which have similar total numbers of mutations assigned to SBS105 its 72% and 32% of the total burden. Surely this is relevant. And why is this?

The total number of mutations found in any given tumor results from diverse mutagenic processes taking place during normal development, aging and tumorigenesis. This combination of processes can be highly variable between different tissues and individuals. The BLCA (bladder) sample with 32% SBS105 contribution (GEL-00089-1) also showed relevant contribution of signatures SBS2, SBS13 and SBS92, as now shown in Supplementary Fig. 1. Firstly, SBS13 and SBS92 are more prevalent among BLCA in comparison to BRCA (breast) (left panel shown below). SBS2 and SBS13 also show a higher mutational burden in BLCA in comparison to BRCA (right panel shown below). Hence, the relative contribution of SBS105 in these two samples can be well explained by the diversity of mutagenic processes involved in different tumor types. Although we believe these differences are relevant, they have been extensively described elsewhere (Degasperi *et al*., 2022, 10.1126/science.abl9283; Alexandrov et al., 2020, 10.1038/s41586-020-1943-3) and are out of the scope of the current work. Secondly, SBS105 relative contribution in the BLCA sample was highest among subclonal variants (~50% contribution for variants with normalized VAF of 0.15-0.35; Supplementary Fig. 4). This exemplifies the complexity of signature admixture in certain samples.



Although Fig 1b addresses this somewhat, the paper is lacking a barplot showing contribution % of these sigs? In all this paper is about 18 patients, and the % mutation burden of these 18 patients should be thoroughly described.

GEL other
SBS95 3/11825 SBS96 7/11825 6 SBS105 2/11825

As suggested to better illustrate the diversity of signatures found in samples with rare CpG mutational signatures, we now include barplots of total exposure (number of substitutions) and relative exposure percentages for these samples in Supplementary Fig. 1.

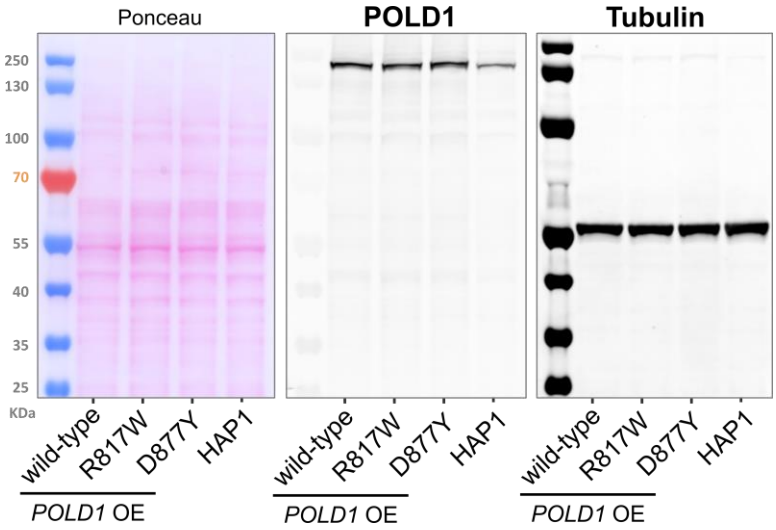
We now include in Supplementary Table 9 the exact number of tumors and patients analyzed from GEL and other sources separately, for each tumor type and predominant CpG mutational signature. This information is now more precisely described in Fig. 1a and the Results section (lines 98-108). Of note, we corrected references to the total number of cases in the GEL series in the Results section (lines 93-100).

In supp table 6 there is a list of supposed variant counts but they all have decimal points? I do not understand how they are not whole numbers. What causes this?
38890,42
35678,37

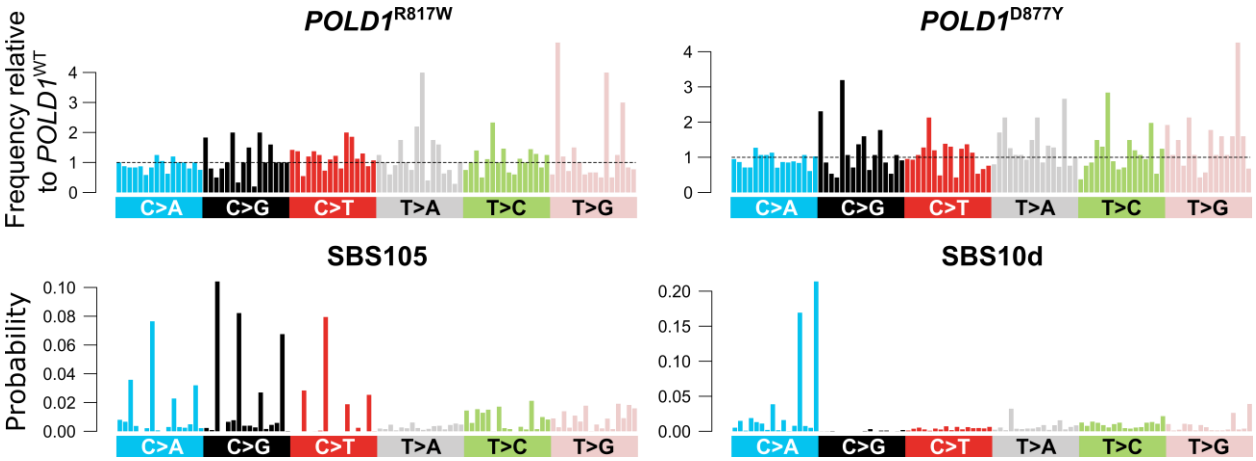
In this work, we rely on a state-of-the-art statistical tool (Degasperi *et al*., 2022, 10.1126/science.abl9283), which is optimized by the use of continuous relaxation for efficiency and simplicity. In practice, signature fitting statistics are computed with substitution probabilities (a continuous variable) that characterize each mutational signature. Mutation counts in each tumor will be equally computed as substitution probabilities, which do not perfectly capture the discrete nature of mutations.

Would be nice to see if you made a POLD1 R817W mutation in a cell line and made a clonal analysis as for the knockout cell lines, would you see the same SBS105 signature after introducing this mutation and clonally expanding a cell?

Following the reviewer's suggestion, we stably overexpressed full-length untagged *POLD1* constructs in HAP1 cells. This was performed with *POLD1* wild-type and polymerase domain mutants R817W (associated with SBS105) and D877Y (associated with SBS10d). Single-cell clones were selected, and each clone showed ~2-fold higher total POLD1 protein levels than parental cells, as confirmed by western blotting (shown below).



Similarly to as performed for *MBD4* and *TDG* knockout models, clones were long-term cultured for 60 days and further subcloned. One subclone per construct was compared to its parental clone by WGS. For unknown reason(s), this experimental model was not able to reproduce either signatures SBS105 or SBS10d (lower panel).



Given that our manuscript mainly focuses on 5mC deamination repair, and due to time constraints, we did not further pursue the development of experimental systems able to reproduce *POLD1* mutational signatures. Nevertheless, we opted to maintain our data on

SBS105 tumors, as it provides valuable information that may guide future research on mutational signatures potentially driven by DNA polymerase.

It is also interesting that you state “Noteworthy, this somatic variant corresponds to a CpG>TpG mutation, which we cannot exclude is a consequence of SBS105 mutagenesis” – a chicken and egg problem. Surely it cannot both be responsible for SBS105 and be caused by it? What else do you suggest as the cause of SBS105 mutations?

CpG>TpG mutations represent the most common substitution class in cancer. Many mutagenic processes could explain the *POLD1*^{R817W} mutation, including SBS105, and we have no means to track back the exact mutational mechanisms involved in each tumor. That this oncogenic mutation was caused by the SBS105 (chicken and egg) is possible, but it would raise many questions that make this hypothesis unlikely.

Is there any insight one can gain from looking at the position of this mutation in the structure?

We thank the reviewer for this suggestion, as it prompted us to compare the tridimensional localization of the different polymerase domain mutations of *POLD1* associated with different mutational signatures (SBS105 or SBS10d). To this goal, we analyze a previously reported structure of the processive human Pol δ holoenzyme (Lancey *et al*., 2020, 10.1038/s41467-020-14898-6). Interestingly, the polymerase domain mutation *POLD1*^{R817W} mutation is proximal to the highly conserved KKRY motif of B-family DNA polymerases (Fig. 2c), which is important for stabilizing the 3′-terminus of the DNA within the polymerase active site and carrying out processive DNA synthesis (Franklin *et al*., 2001, 10.1016/S0092-8674(01)00367-1). Hence, we could speculate that tertiary changes of the KKRY motif by R817W may impair the recognition of misincorporated bases in certain sequence contexts. In contrast, the polymerase domain mutation D877Y is in tridimensional proximity to the iron/sulfur cluster of Pol δ (Fig. 2c), which is essential to its exonucleolytic activity (Jozwiakowski *et al*., 2019, 10.26508/lsa.201900321). Hence, similarly to exonuclease domain mutations, the D877Y mutation may lead to Pol δ proofreading defects. Overall, we show that the tridimensional positions of different *POLD1* polymerase domain mutations are consistent with the mutational signatures with which they are associated. These conclusions have been incorporated into the Results section (lines 219-234).

I feel like SBS95 and the novel signature remain unexplained… When you started the story saying “we next sought to explain the mutations not dependent on methylation” and then you explain SBS105 nicely it would be nice to have more of a section about SBS95 and SBS Novel

We fully agree with the reviewer that we were not able to provide specific mechanisms for SBS95 and SBSnovel. We thoroughly looked for candidate genetic determinants in the 3 tumors with SBS95 signature and found no common variants or pathways that could account for the mutator phenotype. Furthermore, we cannot rule out that this signature is caused by specific/rare environmental factor(s). SBSnovel was found in a single secondary AML sample that was exposed to intensive chemotherapy. This patient harbored a very high TMB, hampering

the identification of gene candidates. We hope that, in the future, additional patients/tumors with SBS95 or SBSnovel signatures will allow us to unravel their causative mechanisms.

Why does the CpA to TpA mutations get classified as SBS96? Shouldn't they be their own signature?

We thank the reviewer for this suggestion, which we strongly considered. However, we followed in this work the policy of the landmark work by Degasperi *et al*., 2022 (10.1126/science.abl9283), which groups tissue-specific mutational signatures into pan-cancer entities. Separating CpG>TpG from CpA>TpA mutations into two distinct signatures would require modifications to their mutational signature discovery tool, which is out of the scope of this work. Most importantly, we show that both mutation classes can be well explained by the same process, unrepaired 5mC deamination due to MBD4 deficiency.

When looking at the CpA methylation in uveal melanocytes what was the extended sequence context? Does it match the DMNT3A motif?

We thank the reviewer for this important suggestion. We now include two new analyses to better characterize non-CpG methylation in normal uveal melanocytes, as well as in three metastatic UVM tumors. First, the genome-wide non-CpG methylation percentages across sequence contexts indicate that uveal melanocytes and UVM preferentially accumulate CpA methylation, more specifically in a CAC context characteristic of DNMT3A-mediated non-CpG methylation (Supplementary Fig. 6). Second, as suggested by the reviewer, we included the extended sequence context around methylated CpAs of uveal melanocytes and UVM. This analysis confirmed the enrichment of the DNMT3A non-CpG methylation motif in methylated CpA sites (Fig. 3g).

For the cut and run, why does the N terminal construct have 10 fold more peaks?

We greatly appreciate the reviewers' questioning of our CUT&RUN data, which has led us to make major adjustments in both our experimental and bioinformatics approaches. We describe these changes in detail below:

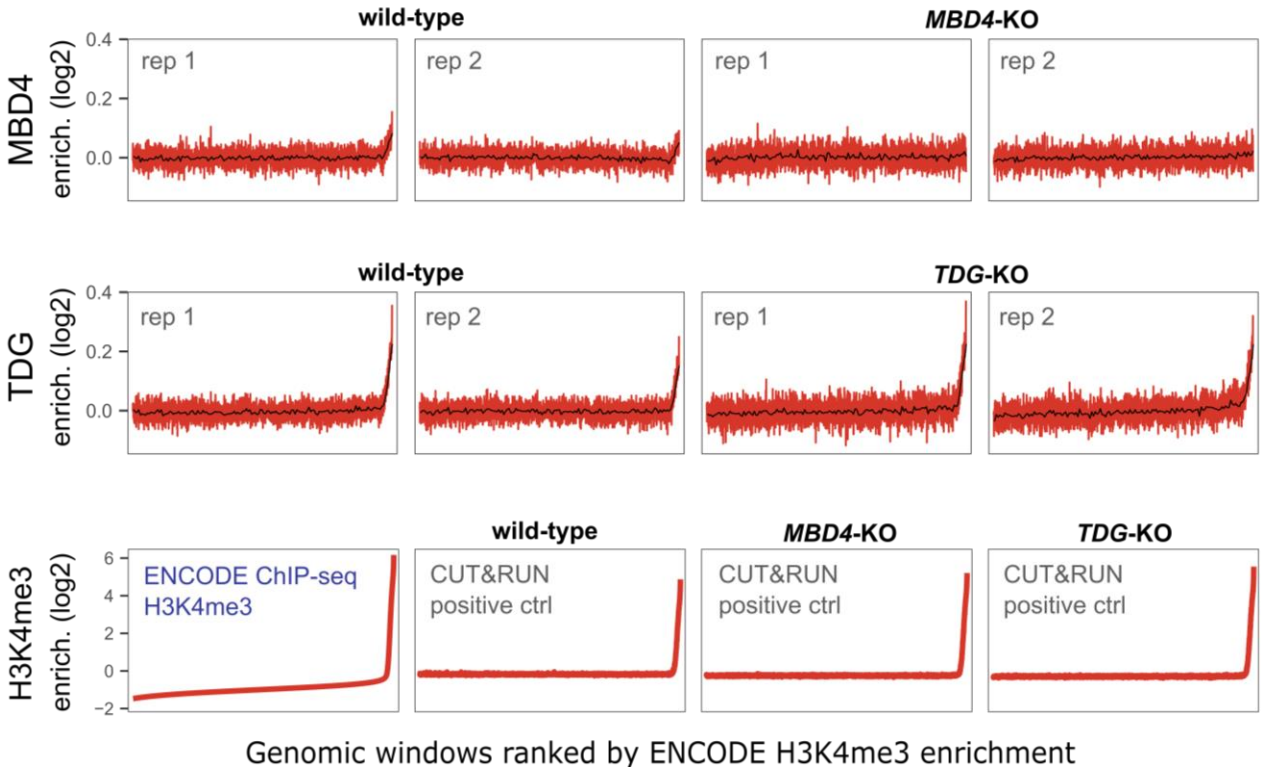
- All CUT&RUN experiments were repeated with an optimized version of the EpiCypher kit (CUTANA ChIC/CUT&RUN Kit V4). With this modification, we observed a better consistency of our replicates due to increased chromatin recovery, and a lower background in open chromatin regions due to a better selection of mononucleosome-sized chromatin fragments before library prep.

- CUT&RUN on tagged MBD4 was originally performed on bulk cell populations expressing highly variable (Supplementary Fig. 9a) and often non-physiological levels of MBD4. We thank the reviewer for bringing to our attention the possibility of degradation fragments in N-terminally tagged MBD4, which was indeed the case for clones with high exogenous expression. We thus selected single-cell clones expressing N- or C-terminally tagged MBD4 at levels comparable to endogenous MBD4, as confirmed by western blotting (Fig. 5a). No major degradation products

of as low as 10 kDa were observed for either clone (Supplementary Fig. 9b). We then obtained FLAG enrichment by CUT&RUN in each tagged MBD4 clone in duplicates, and in parental HAP1 cells as a negative control. Overall, we believe this new experimental design improved our ability to confidently map genome-wide enrichment of tagged MBD4 expressed at biologically relevant levels.

- Baubec *et al.*, 2013 (10.1016/j.cell.2013.03.011) observed that conventional peak calling was largely incompatible with their ChIP-seq data on tagged MBD4 expressed in mouse embryonic stem cells. We indeed observed the same limitation with our newly generated CUT&RUN data. We thus followed Baubec et al's analytic strategy based on genomic windows. We binned the genome into 2-kilobase windows and selected the highest- and lowest-ranked windows based on the enrichment of different histone marks ChIP-seq from ENCODE (Fig. 5b). Tagged MBD4 signal enrichment was positively correlated with activating histone marks H3K4me3 (promoters), H3K27ac (active promoters and enhancers) and H3K4me1 (active and primed enhancers), and to a lesser extent to H3K36me3 (transcribed gene bodies). Conversely, tagged MBD4 signal enrichment was negatively correlated with repressive marks H3K27me3 (polycomb repressed) and H3K9me3 (heterochromatin) (Fig. 5b,c). Overall, we confirm our previous observation that both C- and N-terminally tagged MBD4 preferentially bind to active chromatin domains. In addition, we now show that while tagged MBD4 enrichment is negatively correlated with CpG methylation percentage, it is positively correlated with both CpG and mCpG densities (Supplementary Fig. 10). Hence, our data supports the preferential binding of MBD4 to mCpG-rich genomic regions.

- CUT&RUN on endogenous MBD4 and TDG were repeated using HAP1 cells wild-type and knockout for each gene, in duplicates. We observed higher endogenous MBD4 signal enrichment in H3K4me3 highest-ranked windows (shown below), in agreement with our tagged MBD4 data. This pattern was apparent in both replicates performed in wild-type cells, but not in replicates performed in *MBD4* knockout cells. However, this experiment suffered from poor sensitivity, and we believe it would add little value to our revised manuscript. The experiment on endogenous TDG showed poor specificity instead, with similarly high TDG enrichment in H3K4me3 highest-ranked windows in wild-type and *TDG* knockout cells (shown below). Notably, H3K4me3 positive control reactions performed in cells of the three genotypes showed equally strong signal enrichment in ENCODE ChIP-seq H3K4me3 highest-ranked windows. We believe that the limited sensitivity and/or specificity of these experiments were largely dependent on the quality of the antibodies currently available. Hence, we opted to withdraw CUT&RUN data on endogenous MBD4 and TDG from our manuscript. Despite this technical limitation, we believe that the overall conclusions of our study remain largely unchanged. Firstly, we show that both N- and C-terminally tagged MBD4 preferentially bind to active chromatin domains, in agreement with the literature (Baubec *et al.*, 2013, 10.1016/j.cell.2013.03.011). Secondly, it has been previously shown that tagged TDG preferentially binds to active chromatin domains rich in H3K4me3 in mouse embryonic stem cells (Neri *et al.*, 2015, 10.1016/j.celrep.2015.01.008).



How do you explain that MDB4 deficiency is responsible for 5mC mutations when MBD4 doesn't seem to bind to methylated DNA? What is the mechanism here? It is clear that knockout of MBD4 increases mutations as well, but if the cut and run doesn't show binding it is a bit strange. A comment about this would be nice.

CUT&RUN doesn't provide enough resolution to precisely identify whether MBD4 is bound to localized methylated CpGs in an otherwise low methylation region. Moreover, the glycosylase activity of MBD4 is largely independent of the methylation state in the vicinity (Ishibashi *et al.*, 2008, 10.1128/MCB.00588-08). In the revised version of this manuscript, we now show that while tagged MBD4 enrichment is negatively correlated with CpG methylation percentage, it is positively correlated with both CpG and mCpG densities (Supplementary Fig. 10). A similar pattern was previously described for biotin-tagged MBD4 on mouse embryonic stem cells (Baubec *et al.*, 2013, 10.1016/j.cell.2013.03.011). Hence, our data supports the preferential binding of MBD4 to mCpG-rich genomic regions.

Minor comments:

Clarify patients, in the methods its says 25 patients but in supp table there are only 6 patients not in GEL. Why weren't the others included in the WGS cohort? When you include different patients for different experiments its hard to understand. It would be nice if there was all experiment types for less patients rather than different patients for different experiments, particularly when you are looking at rare phenomenon.

Following reviewer's remarks, detailed information on the sequencing modality used for each in-house sample is provided in Supplementary Table 1, which lists samples analyzed by WGS, WES and/or RNAseq. This information was included in the Methods section (lines 467-471). Unfortunately, analyzing each sample by all three modalities was not possible. Nevertheless, the extended series of UVM cases analyzed exclusively by WES and RNAseq allowed us to pinpoint the spectrum of CpG>TpG driver mutations specifically linked to MBD4 deficiency. This would not have been possible had we restricted our analysis to samples with WGS data.

Mutational signature analysis was performed on filtered somatic variants obtained by WGS only (2 in-house *MBD4*def tumors, 4 publicly available *MBD4*def tumors, and 12,726 tumors from GEL). This information is now included in the Methods section (lines 89-93). Importantly, the sheer number of mutations obtained by WGS in a small number of samples still allowed us to comprehensively dissect the genome-wide patterns of mutagenesis caused by 5mC deamination in different tumor types.

Make data available, it says its at EGA but what is the accession number?

Newly generated WGBS, WGS, WES and RNAseq data from Institut Curie patients are now deposited in the European Genome-phenome Archive (EGA) database under accession EGAS50000000536. Newly generated patient sequencing data from collaborating institutions will be made available on request. CUT&RUN data are deposited in Gene Expression Omnibus (GEO) under accession GSE275181. This information has been included in the Data Availability section of the manuscript (lines 779-798).

Fig 1 legend should be 5mC

Fig. 1 legend was corrected.

Fig 3C. Is it correct to adjust the scales between what should be similar results. E.g the N term and C term overexpression constructs a shown to be similar but the scale is adjusted to show this. What is your reasoning for adjusting the scales so they seem similar?

As mentioned above, our new CUT&RUN experiment was performed with carefully selected single-cell clones expressing exogenous tagged MBD4 at levels comparable to endogenous MBD4. As now shown in Fig. 5b, FLAG enrichment over IgG had a similar range for both N- and C-terminally tagged MBD4, and no scale adjustment was necessary.

Rebuttal letter

The authors greatly thank the reviewers for their additional comments on the revised version of our manuscript. We believe that the changes prompted by this review considerably improved the overall quality of our manuscript.

In this revised version, changes to the main text have been marked in blue color. Figure 2a had been generated with a subset of variants of each signature. This has been corrected, and the interpretation of this figure remains unchanged. We made corrections to the Methods section regarding the filtering criteria used for the calculation of mutation rates (lines 694-697) and parameters used for the calculation of local sequence context statistics (line 710). The interpretation of these analyses remains unchanged. We also included missing information on the snRNAseq dataset used (lines 698-700).

Additional changes have been made following the Journal's editorial policy. Results headings have been adapted to contain up to 60 characters. Main figure legends have been corrected to follow the editorial checklist and to contain up to 350 words (without titles). A missing reference was added, and the final References list has been adapted to contain exactly 70 entries.

Below, we provide point-by-point responses to comments by reviewers 1 and 3.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Dear Authors
I honestly and greatly appreciate the extensive work you undertook in a rather short time to answer and elaborate on the reviewers and especially my many, and often admittedly tough, interrogation. I want to apologize for my oversight concerning part of the WGBS methodology and thank for the individual explanations delivered in the rebuttal.
The authors addressed all of my concerns and the added results and controls greatly increase the confidence in the data to support the conclusions. Furthermore, do the presented results not "just" shed light on the clinical observation, but deliver thorough data pointing to new directions and hypotheses that will surely inspire future research about deamination/MBD4-independent C>T transitions and the dynamic (epi)genome.

A few minor points:

The authors write that the MBD4-FLAG proteins are expressed at a comparable level as the endogenous MBD4. The western blot in Fig. S9, clearly shows higher levels of the FLAG-MBD4, which corresponds to about double the level of endogenous protein expression when roughly measured by densitometry and normalized to the ponceau signal. It is a little deceiving – densitometry could be added to the WB. Nonetheless, I don't question the later obtained results.

We agree with the reviewer that the previous wording was slightly misleading. Densitometry analysis was performed and added to Supplementary Fig. 9c. The text of the Results section was adapted accordingly (lines 355-356).

The authors present a WB analysis for MLH1 (Fig. 5a, 6a), which is not mentioned in the text. I understand this is to control for potential confounding effects of MMR, it should probably be very briefly mentioned.

We thank the reviewer for this suggestion. We now state in the Methods section (lines 792-793) that western blotting of MLH1 was used to control for potential confounding effects of MMR.

The authors repeated CUT&RUN assays to control for potential artefact observations, amongst others, due to antibody unspecificity. Unfortunately, this was indeed the case but this observation is critical and commendable. I encourage the authors to maybe include a statement about the (current) infeasibility of CUT&RUN with the used antibodies (e.g. in the methods).

We appreciate the reviewer's comments. We now state in the Methods section (lines 760-762) that our CUT&RUN attempts with currently available anti-MBD4 antibodies showed poor sensitivity and/or specificity.

The authors state that "Notably, while tagged MBD4 enrichment was negatively correlated with CpG methylation percentage" (Fig S10a).
i) could you maybe overlay the plot with the MBD4-signal on a second y-axis, to give a better grasp on how MBD4 binding is distributed over the inspected windows (i.e. is it gradual or not?/ does correspond "lowest" to no enrichment..)

As suggested by the reviewer, we now include FLAG enrichment values for the ranked genomic windows presented in Supplementary Fig 10a. We believe the interpretation of this figure remains unchanged.

ii) This is rather surprising, given its role as mCpG binding protein and the contrast to the currently accepted view that “highly” or rather more stably methylated CpGs are more prone to spontaneous deamination. Including the above-mentioned data might relativize this apparent discrepancy. If not, I would be looking forward to a short comment about that.

Sincerely, Simon Schwarz, DBM University of Basel

Although we agree that the negative correlation between tagged MBD4 enrichment and mCpG percentage is counterintuitive, our data confirms that MBD4 preferentially binds to genomic regions with higher mCpG density. This is in agreement with the current literature (Baubec *et al*., 2013; 10.1016/j.cell.2013.03.011). Of note, global CpG and mCpG densities are highly correlated genome-wide, partially explaining this surprising pattern. In addition, MBD4 has been shown to recognize a wide range of 5-methylcytosine deamination and oxidation derivatives (Otani *et al*., 2013; 10.1074/jbc.M112.431098), potentially contributing to its genomic distribution.

Reviewer #3 (Remarks to the Author):

This revised manuscript is much improved. The authors have responded to the identified flaws quite well. I am pleased the flaws in the CUT&RUN were corrected and the decision to remove endogenous results also makes sense. I hope to see this amended and published in a later manuscript. I also really liked that the extended context of the CpA mutations matched the DNMT3a motif nicely.

I am happy to accept this manuscript after the resolution of the below comments:

Figures need to define all statistics, what do the box and whiskers of the boxplots represent (median 25th 75th percentile?)
what is the shaded region in scatterplots eg 1e

We thank the reviewer for noticing the missing information. Boxes indicate the median, 25th and 75th percentiles. Whiskers extend to the largest or lowest value up to 1.5 times the distance between the 25th and 75th percentiles. This information has been included in all the concerned figure legends. The shaded region in Fig. 1e represents the 95% confidence interval. This information has been included in the figure legend.

It was commendable that you attempted to recreate the R817W mutation in POLD1 but didn't see any effect. Do you think this could be because there remains sufficient levels of the endogenous wildtype enzyme?

Considering that both SBS105 cases harbored heterozygous *POLD1*^{R817W}, it is unlikely that residual wild-type POLD1 activity could explain the absence of an observable phenotype in our cell model. However, we can point out a few caveats related to our experimental approach. First, we could not fully account for the diversity of *POLD1* isoforms. Notably, a detailed description of the respective roles of the different human *POLD1* isoforms is lacking in the literature. Second, our approach did not capture the gene expression dynamics of *POLD1* through the cell cycle, which may be relevant. Hence, we believe that base editing of the endogenous *POLD1* locus may be required in future experiments. In addition, we cannot exclude limitations inherent to *in vitro* modeling.

Code availability: Great to see the data deposited, but the analysis involved a lot of code? Shouldn't this also be included to enable recreation of the results? Also source data.

In this study, we did not develop custom bioinformatic tools. Instead, we applied standard and published bioinformatic resources, as described in detail in the Methods section. All publicly available data used are listed in the Methods and Data Availability sections. Source data file is now provided for all plotted Figures, except for those whose data are already provided as Supplementary Data.

Fig 5d Melanocyte SBS1 is blank

SBS1 was not found at a sufficiently high percent contribution in any of the analyzed Uveal Melanoma cases, corresponding to the melanocyte lineage presented in our manuscript. Hence, the data is not available for Fig. 5d. Of note, the numbers of samples included for each tumor type and lineage are listed in Supplementary Table 9.