***Response to reviewers***

Manuscript JoVE52246R2

'Enhanced Reduced Representation Bisulfite Sequencing for assessment of DNA methylation at base pair resolution'

We appreciate the opportunity to revise our manuscript and are hereby submitting revised documents along with responses to reviewers. Editorial and reviewer comments are in italic bold font below and responses to comments follow each respectively.

***1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.***

Thank you for the opportunity to review the manuscript for any errors. All authors have thoroughly proofread the manuscript.

***2. JoVE cannot publish manuscripts containing commercial language. This includes trademark symbols (™), registered symbols (®), and company names before an instrument or reagent. Please remove all commercial language from your manuscript and use generic terms instead. All commercial products should be sufficiently referenced in the Table of Materials and Reagents.***

***For example: All Cells, LLC, etc.***

We recognize JoVE’s restrictions on the publication of commercial language and the AllCells terminology has been removed

***3. Please double-check the in text references to Protocol steps for accuracy.***

Thank you for the opportunity to review the manuscript to ensure reference list accuracy. All authors have reviewed this information

***4. Please place all URLs in the Table of Materials and Equipment.***

All pertinent URLs have been included in the Table of Materials and Equipment. For large equipment that is commonly used in laboratories, no specific URL was included.

***5. Please include all relevant details that are required to perform the step in the highlighting. For example: If step 2.5 is highlighted for filming and the details of how to perform the step are given in steps 2.5.1 and 2.5.2, then the sub-steps where the details are provided must be highlighted.***

***6. Please highlight 2.75 pages or less of the Protocol (including headings and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. The highlighted steps should form a cohesive narrative with a logical flow from one highlighted step to the next. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader.***

Thank you for the opportunity to review the manuscript to ensure all protocol steps are properly highlighted. Revisions to the sections annotated for the potential video have been included in the updated version.

***Reviewers' comments:***

***Reviewer #1:***

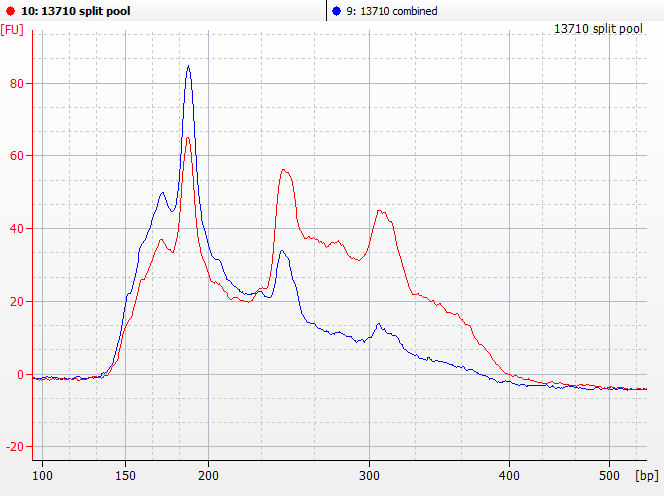
***This article titled "Enhanced Reduced Representation Bisulfite Sequencing for assessment of DNA methylation at base pair resolution" is a clearly written article that describes the steps involved in generating the high-throughput sequencing libraries for the researchers interested in cytosine methylation patterns at CpG rich regions at a reduced cost. The steps of the protocol for ERRBS are written with clarity giving details about the reagents and the alternatives available. This article is well within the scope of the journal and the details are sufficient for this type of article. In this article, the authors have started by introducing the role of cytosine methylation in the genome, methods in use to analyze the cytosine methylation and the advantage of ERRBS over the available methods. All the steps in the protocol were explained with enough details that will enable reproducibility. Authors have also extended the article by briefing about the methods used to analyze the generated data. Figures are representative, simple and of good quality. Figure legends explain the figures very well.***

We thank reviewer one for his/her thorough input. Below are comments to address each point raised.

***However, there are a few points to clarify and typos to correct:***

***All through the protocol, the authors have treated the two fragments sizes (135-240 and 240-410 bp) separately and subjected them to similar reactions and treatments. Finally they mixed the two fragment sizes. Authors may clarify why they were treated the fragment sizes separately and not as a single fraction (135-410). What is the advantage in treating them separately while authors are not excluding any intermediate fragment sizes? Is this specific to Pippin Prep?***

The rationale for the division of the library product into two fragments is to ensure efficient library product amplification during the PCR step (protocol number 7). Splitting the fractions for amplification and then combining them in equimolar amounts to make the final library for sequencing results in a more equal distribution of fragments across the entire size range. The bioanalyzer traces below illustrate the difference between a library created from a single 150-400 bp fraction (blue trace) and a library created from two fractions (red trace). The library created from a single fraction skews toward smaller fragments while the two-fraction library has a more even distribution of fragments.



Base pairs

Fluorescence units

***It is likely beneficial to the users of this protocol to have a manual gel-based size selection protocol, as many (most) may not have access to a Pippin Prep. Although one is referenced, this would provide a single protocol for ERRBS.***

We thank the reviewer for this recommendation. Protocol steps for a manual gel-based size selection have been included in the revised manuscript (protocol section 5.2).

***Lines 538-543 give a good description of sequenced read depth to mapped reads to # of CpGs obtained. However, the authors could aid the readers/users by providing a barchart/histogram/graph that shows these numbers in 10 or 20 million sequenced read increments; i.e. for each read depth increment how many reads mapped, and how many CpGs are recovered or gained by going deeper. This will allow the user to judge how deep to sequence. The example given is for a full lane, which seems rare for RRBS experiments. If a full lane isn't needed for ERRBS, as is the case for RRBS, then multiplexing should be noted.***

We thank the reviewer for this recommendation. We have added summarized results from multiplexed sequencing runs in the manuscript. We have added a figure (Figure 5) representing this information in comparison to expected data from sequencing a single ERRBS library per lane and simulated downsampling results from three representative sequencing runs (sampling 5 times for each comparison). We have also added Table 4, which includes average results from multiplexing ERRBS libraries that we have performed in the past year. Commentary about the expected results in also included in the results section.

***The authors may consider discussing alternatively spiking in sonicated lambda DNA for the bisulfite conversion efficiency with the caveat that it is only detected after sequencing the library not prior.***

We thank the reviewer for this recommendation. We agree that such a control is beneficial to include with the disadvantage of not being able to determine conversion efficiency until after sequencing. Should the conversion efficiency be suboptimal, this process would pose a high financial cost and we recommend the use of a parallel control that can be independently validated as described in the protocol. We have revised the manuscript to include this information for user’s choice as a note in protocol section 6 and have referenced a publication which describes such an approach1.

***At lines 465-474, the authors have specified the usage of "FAR" software referring to an article about "Flexbar" software (Ref 28) for trimming the reads. I noticed that "FAR" is now defunct and not available for downloads. Authors should clarify the name of the software and its reference (if they wish to name it) and its reference.***

The reviewer is correct. Flexbar is the current name of the FAR software, and we have updated the text (protocol step 11.3) to reflect this information inclusive of the current reference for this tool. Any software tool that removes adapter sequences could be used (cutadapt, trimmomatic, etc.) in this step of the analysis and we have included additional examples in the manuscript for the readers’ consideration.

***At lines 493-514, Bismark is a good tool but is under continuous development with additional scripts available for format conversion and also enables usage of multiple aligners (Bowtie and Bowtie 2 for ungapped and gapped alignments). So, either authors may update the text to refer Bismark's perl scripts or generalize the word "PERL script" to "custom scripts".***

We thank the reviewer for this suggestion. We have updated the text to recommend the use of custom scripts when appropriate.

***Figure 5D is misleading without contrasting the R squared value of different biological samples. RRBS enriches for CGs where the majority are largely 100% methylated or non-methylated; therefore it is likely that even two different cell types will have an equally high correlation.***

We thank the reviewer for this comment. We have performed correlation analyses of CpGs covered and their methylation levels on various sample types and in all cases the R2 values were below that reported for the technical replica included in this manuscript. We have included a few examples below:

|  |  |  |
| --- | --- | --- |
| **Sample 1** | **Sample 2** | **R^2 value** |
| IMR90 | HCT116 | 0.4675 |
| AML3 | CB | 0.8385 |
| AML3 | NB | 0.8416 |
| AML1 | AML2 | 0.7665 |
| AML1 | AML3 | 0.8647 |
| AML2 | AML3 | 0.7686 |

IMR90 = normal human lung fibroblast cell line

HCT116 = human colorectal cancer cell line

AML = Acute Myeloid Leukemia samples (1-3)

CB = centrocyte B cells isolated from tonsils

NB = Naïve B cells isolated from tonsils

We have modified the text to indicate R2 value ranges expected between technical versus biological samples as above-noted. Should the reviewers deem additional figures necessary, we can include those as well.

***@line 164-168: pH of the Tris-Cl is missing. Since pH of the buffer is crucial and determines the phase separation of DNA and RNA, it is worth mentioning the pH of the buffer.***

Thank you for the reviewer’s attention to this omission. We agree that buffer pH’s are essential to successful protocol implementation and the information has been included in the revised manuscript.

***@line 167: following Tris- Ethylenediaminetetraacetic acid “(TE)” could be shown***

We have included this edit in our updated manuscript

***@line 178: Alcohol precipitation is a method that depends on the concentration of the alcohol and sodium acetate; therefore, the authors may consider rephrasing the amount of alcohol used (750 microlitres be changed to 2.5 volumes).***

Thank you for the recommendation; we have revised the protocol to ensure clarity and have added volumes to the solution quantities already noted.

***@line 244: "unicorporated" is misspelled.***

Thank you for the reviewer’s comment. We have revised the manuscript accordingly

***@line 511-514: The recent versions of Bismark generate the SAM formatted output by default. No need of any further scripts to process Bismark's output. Authors may generalize about the possible ways to visualize the methylation data (there are other alternative ways to visualize the data in addition to wiggle format).***

We thank the reviewer for this comment. We use a previous version of Bismark that does not generate SAM formatted output but instead provides a bismark2SAM\_v5\_xm.pl to convert Bismark alignments to SAM format which we then convert to BAM format using SAMTOOLS. In the updated manuscript, we have included additional viewing options to wiggle track formats, including bedGraph format, the UCSC genome browser and IGV.

***Reviewer #2:***

***Garrett- Bakelman et al have described a high resolution genomic methylation analysis platform (e-RRBS) protocol that will benefit epigenetic researchers especially those studying primary tissue samples with low input DNA.***

***The article is well written and organized and the companion video will help enhance the visualization of the protocol.***

We thank reviewer two for his/her thorough input. Below are comments to address each point raised.

*Major concerns:*

***1. Steps 10 and 11 have not been included in the visualization - these will be important especially to convey key aspects of the data analysis (step 11)***

We thank the reviewer for this recommendation. Due to the length of the protocol and JoVE’s guidelines for highlighted segment inclusions, we will be unable to incorporate step 10 and 11 into the visualization process. We agree that these steps are essential to successful implementation of the protocol and hence have added additional information in them. We have revised our manuscript to include further details on the library pooling protocol (protocol step 10). Furthermore, we have expanded our supplemental code file to include a full list of detailed commands for data alignment.

*Minor concerns:*

***1. Use of a high complexity library lane as a control for cluster density during sequencing is mentioned on page 15. Can the authors add 1-2 sentences on how this is used?***

We thank the reviewer for this recommendation and we agree that further explanation would be helpful. We have modified the manuscript in the discussion to clarify the recommendation to use a high complexity library lane (PhiX control or other library type) as a dedicated independent control lane. When designated as a control lane for the respective sequencing run, it can serve as the basis for the matrix generation which is utilized during the first four bases of sequencing to detect cluster positions. The higher quality reads captured will raise the mean coverage per CpG site by 5.2 (n=4).

***2. Add in discussion whether the technique can be adapted for multiplexing?***

We thank the reviewer for the recommendation. We have expanded the manuscript to include further detail on how the protocol could be adapted for multiplexing (protocol step 12). We have also added figure 5 which highlights expected outcomes from standard and multiplexed approaches.

***Reviewer #3:***

***Editor’s Note:*** *Please note that the reviewers raised some significant concerns regarding your method and your manuscript. Please thoroughly address each concern by revising the manuscript or addressing the comment in your rebuttal letter.*

*Manuscript Summary:*

***This protocol describes what the authors consider an improvement of the original RRBS protocol and which they called "enhanced RRBS". However, the changes compared to the original protocol are incremental in nature and do not justify the new name, especially because the improvements in coverage come at the cost of significantly increased costs per sample and therefore constitute a different trade-off rather than a dominant advance.***

***Furthermore, the comparison between the original RRBS protocol and the "enhanced RRBS" protocol is no longer relevant because the original RRBS protocol has been superseded by much-improved versions in many labs, which provide various improvements that the described "enhanced RRBS" protocol is lacking.***

We thank reviewer three for a thorough review of our manuscript. The manuscript submitted details the current protocol we use based on our previous publication (Akalin and Garrett-Bakelman, et al. PLoS Genetics **8**(6) in 2012) with the goal of providing guidance and sufficient detail for setting up and implementing the protocol in a laboratory novice to this technique. We have addressed each concern raised below.

***Compared to the current state-of-the-art, the most severe flaws of the presented protocol are as follows:***

***1. Low throughput: The protocol contains several steps that are inherently low throughput, which makes it impossible to effectively process large cohorts****.*

We agree that the ERRBS protocol as described is limited to the preparation of 10 - 15 samples per week by a single individual. In our experience, the vast majority of projects profiled include up to 30 samples. The protocol could be converted to a high throughput protocol if the user or lab implementing it has an automated next generation library preparation workstation. However we do not think this is feasible in most labs and we have used the described approach successfully and reproducibly in over 2000 samples.

***2. High cost: The protocol is clearly not optimized for cost and lacks even basic features such as support for sample barcoding and multiplex sequencing.***

We thank the reviewer for this comment. We routinely perform our protocol to acquire data with high coverage per CpG site in order to perform analysis that requires high coverage depth (for example to detect epigenetic polymorphisms as described in Landan, et al.2). While we agree that barcoding and multiplexing ERRBS will yield a lower cost per sample protocol, this will result in a reduction in the number of CpGs covered and the coverage at each CpG site represented in the data. We have revised our manuscript to include further details on how the protocol can be modified for a multiplexed sequencing approach (protocol step 12) and have included data on the number of CpGs and depth of coverage anticipated from such an approach for the readers’ consideration (Figure 5 and table 4).

***3. Infeasible for most labs: The protocol uses specialized equipment for size selection that is not only a severe throughput bottleneck but also unavailable in most labs.***

We thank the reviewer for this comment. We have incorporated a section in the protocol describing a gel extraction approach for size selection, which should be feasible in all laboratories (step 5.2). Unfortunately unless the protocol is automated, this step is a bottleneck, as it is in many other next generation sequencing library preparation protocols.

***4. No proof of robustness: I would expect that the authors summarize their experience with the protocol for at least a couple of dozen (or better: hundreds of) samples run in their lab over the last years, such that the reader gets an impression of the protocol's robustness.***

We thank the reviewer for his/her comment about protocol robustness. We had indicated in the representative results section of the manuscript our experience in terms of data quality and quantity from human samples sequenced at 50bp single end reads in the past few years (cluster densities, pass filter rates, number of uniquely aligned reads per lane, alignment efficiency, the number of CpGs covered per sample and the average coverage at each CpG covered). Should there be any additional information the reviewer deems appropriate for inclusion, we would be happy to do so.

*Additional Concerns:*

***1. The protocol contains far too many cleanup steps that require a large amount of input DNA. Although the authors claim that the procedure would work even down to 5ng of input DNA they never actually show data that support this claim.***

We thank the reviewer for this comment. In our PLoS Genetics paper (Akalin and Garrett-Bakelman, et al. 2012), we had performed a titration of input DNA (5, 50 and 1000ng) which revealed high reproducibility of our current ERRBS protocol when data was examined for all CpGs covered as well as CpGs found in CpG islands and promoters (pearson correlation between technical replicas of ERRBS using 5, 50 or 1000 ng genomic DNA from the HCT116 cell line). This has been referenced in this manuscript.



***2. The described bisulfite conversion protocol has been superseded by newer and more efficient protocols.***

We agree that several newer kits for bisulfite conversion have been brought to the market however we have found that the kit we use is highly reproducible. In our hands, the bisulfite conversion rates observed are consistently higher than those reported to be guaranteed by kits offered by Zymo research and Epigentek (99 – 99.5%), with an average conversion rate from 390 sequencing runs of 99.85 +/-0.04 %. We have added alternative reagents to the Table of Materials and Equipment which may be used should the user prefer.

***3. The bisulfite conversion controls are added too late. Why don’t the authors spike them in at the very beginning (for end repair, A-tailing and ligation)? This would enable them to control individual molecules on the HiSeq sequencer and not just a mixture of individual signals on the Sanger sequencer. Furthermore, no over-conversion controls are used.***

We agree that a spike-in control could be beneficial to include. However a disadvantage compared to the control detailed in the protocol is not being able to determine conversion efficiency until after sequencing. Should the conversion efficiency be suboptimal, this process would result in low quality data due to inefficient conversion and pose a high financial cost that could have been avoided. We recommend the use of a parallel control that can be independently validated as described. We have revised the manuscript to include the option of a spike-in control for the user’s choice (Note in protocol step 6). While we do not routinely test for over conversion, thorough review of this concern with the research division at ZymoResearch’s revealed that over-conversion is estimated to be sub-percent and that the impact of over-conversion on data quality would not be significant unless the bisulfite treatment would be extended for an additional four hours than currently used. Finally, review of select papers published about RRBS and other related protocols does not reveal routine use of an over-conversion control in published protocols to date (for example 1,3-5).

***4. Running 18 cycles by default can lead to an amplification bias in some cases. A qPCR should be performed first, to calculate the ideal cycle number for enrichment.***

We agree that a qPCR step for library quantitation prior to PCR amplification could be considered for inclusion in the protocol, however we routinely do not implement it. Based on our assessments, introducing a qPCR step will not result in a change in the protocol proposed and 18 cycles of PCR yield sufficient material for sequencing from 5-50ng starting materials (typical yields are: 25 µl of 30 nM for lower fraction and 25µl of 8 nM for the higher fraction from 5ng starting materials and 50 µl of 300 nM for lower fraction and 50 µl of 20 nM for higher fraction from 50 ng starting materials). This protocol is continuously under improvement and qPCR quantitation is one aspect we are exploring for further development. We have included a comment in the discussion raising the possibility of using qPCR for further PCR optimization.

***5. For a technique like RRBS where there is no possibility to discriminate for PCR duplicates, 18 PCR cycles are already the upper limit that should be applied to a library during enrichment. I don't see how an efficient library should be generated with 5ng of starting material when there is the need for 18 cycles of PCR already with 50ng of starting material.***

We have used differing inputs for the same samples and we have not observed higher rates of clonality at MspI sites or a lower correlation (as highlighted under comment number one). We agree that lower PCR cycle numbers are advantageous, but the distribution of covered sites does not indicate an issue with over-amplification except for rare loci with extreme coverage results (see figure 5B in the manuscript). We have successfully prepared and sequenced ERRBS libraries with 15 cycles of PCR starting with 50 ng of DNA. We have included a notation in the protocol that advises that PCR cycle reduction is possible as further optimization of the protocol (protocol step 7). We recommend that novice users start with 18 cycles of PCR to ensure success with the protocol. Experienced users can reduce the number of PCR cycles and still achieve sufficient library product for sequencing. We have indicated that titration of the number of PCR cycles is possible in the paper for the readers’ consideration. Finally, we have found that 18 cycles of PCR are necessary when starting with 5 ng of material.

***6. Throughout the protocol, large reaction volumes and excessive amounts of enzymes are used, which makes the protocol unnecessarily expensive.***

We appreciate this comment. The protocol submitted is the current approach used. The cost of the reagents and consumables used for a single ERRBS prep is $120.00. Compared to other next generation sequencing protocol kits used, such as RNAseq and exome capture, we do not find that the cost of the library preparation is excessive. The protocol is continuously undergoing improvements and this is an aspect that is assessed on an ongoing basis.

***Reviewer #4:***

***The manuscript describes improved RRBS procedure for genome scale DNA methylation analysis.***

***Whole genome bisulfite sequencing is the ultimate method for studying DNA methylation genome wide. RRBS focuses on most relevant genomic regions at an affordable cost. ERRBS presented here is very similar to RRBS but with modifications which improve reproduciblity and coverage of additional relevant genomic regions.***

***The protocol is described in detail in a user friendly way, paying attention to key considerations. Having a detailed protocol with video demonstrations and a complementary analysis guide is very useful for someone trying to setup this assay.***

***With this objective in mind I list some minor comments and some cases that additional orientation and tips can prove useful when setting up the protocol.***

We thank reviewer four for his/her thorough input. Below are comments to address each point raised.

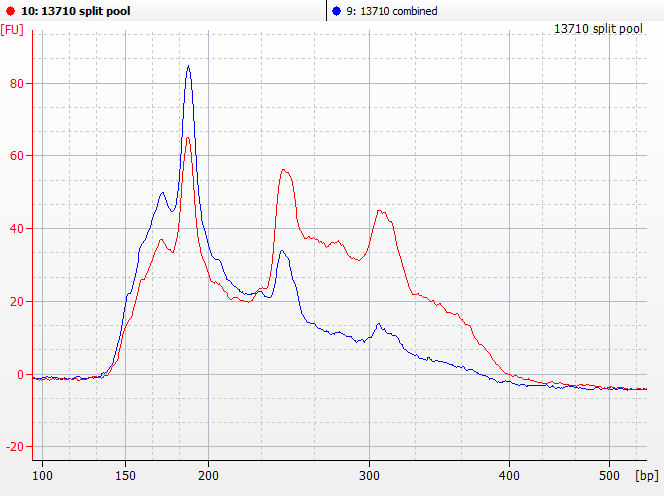
***comments to authors***

***1. The protocol is described for 50ng DNA. If the authors have experimented with higher (or lower) amounts of input is will be helpful to provide some guidelines as to how this affects on the protocol downstream (for example on adaptor concentrations, size selection step, number of PCR cycles).***

We thank the reviewer for this recommendation. We have revised the manuscript to include Table 3, which details how we recommend changing the protocol for various starting material quantities. We agree that this information will be useful to readers

***2. The protocol is described for a single sample (which after size separation becomes two samples which are processed separately). It is common practice to pool samples together after adaptor ligation for subsequent processing. Instead, the authors choose to split each fraction into 4 amplification tubes. It will be interesting to know what led to this decision. In this case, I assume that the two size ranges have to be processed separately at least until amplification. But in theory, when several DNA samples are studied, it may be possible to pool samples together after size selection so that the high fragment size fraction from several samples are bisulfite converted and amplified together. This should reduce costs and reduce heterogeneity between samples introduced by processing. Assuming the authors considered this possibility, it will be helpful if they can share their conclusions, and if this possibility is practical, offer some guidelines (or full details) how can this be done.***

We thank the reviewer for this comment. The rationale for the division of the two fragments is for the optimization of the downstream PCR amplification step. We have found that maintaining independent higher and lower fractions results in higher efficiency of PCR amplification and more equal representation of each in the final library pool (please see figure below: red trace represents a pooled ERRBS library for sequencing prepared per our protocol; blue trace represents a library prepared as a single fraction isolated from size selection).



Base pairs

Fluorescence units

We agree that pooling lower and higher fragments from several samples together for sequencing can be cost effective, however only if limited coverage per CpG is a desired result. Furthermore, any pooling approach after size selection will be limited by the optimal input DNA quantity that the bisulfite conversion method used can utilize. The protocol submitted for publication is our current version in use. It is continuously under development and this recommended consideration for change in the protocol can be taken into consideration for further improvement. For the purposes of performing analysis to determine epigenetic polymorphisms as described by Landan et al.2, it is necessary to obtain high coverage data and we recommend the use of the equivalent of single lanes per sample for such data acquisition.

***The protocol describes pooling of the two fractions (high and low size) for one sample but does not address pooling of different libraries (different DNA sources) for running on one sequencing lane.***

We thank the reviewer for this comment. While more cost effective, multiplexing several samples per lane will significantly reduce the number of CpGs covered and the depth of coverage for each CpG represented in the data. We have revised the manuscript to include additional information on a multiplexing approach that can be used should that be a desired approach (protocol step 12), as well as information on data yields possible from such an approach compared to sequencing a single ERRBS library per lane (Figure 5 and Table 4).

***3. The protocol emphasizes to importance of sequencing control unconverted sample (such as RNAseq ChIPseq) in parallel to bisulfite to ensure correct base calling. The term control lane suggests it is a different lane. The authors should clarify whether they mean that a control sample can be multiplexed with the bisulphite converted sample/s and run in the same lane.***

We thank the reviewer for this comment and agree that further clarification could be useful to the readers. We have added additional details in protocol step 10 to address this topic, including the clarification that we are recommending the use of an independent lane for a control sample.

***I could not find here a comment on the desired number of reads for the human/mouse genome, or reference to more elaborate examination of this issue.***

We thank the reviewer for his/her comment and agree that such information would be helpful to the readers at large. We have included a figure with simulated estimates of CpGs covered per read counts for human genomes compared to sequencing a full lane per sample to address this question. Furthermore, we included data from our experience with multiplexed sequencing for the readers’ reference (Figure 5 and Table 4).

***Figure 3: images for good and bad preparations are presented. It will be useful to provide more guidance as to what to pay attention when assessing the outcome.***

We thank the reviewer for this comment. We have added figure 3G to further visualize high quality library products appropriate for sequencing. We expanded the discussion to address key features of library quality control and potential failure criteria to consider. Furthermore, in the results, we have included assessment data from ERRBS sequencing runs (bisulfite conversion rates, cluster densities, pass filter rates, number of uniquely aligned reads per lane, alignment efficiency, the number of CpGs covered per sample and the average coverage at each CpG covered). Should there be any additional information the reviewer deems appropriate for inclusion, we would be happy to do so.

*Minor Concerns:*

***1. Specify the PH when buffers are mentioned (for example item 1.4.1.1 - line 164)***

We have included this edit in our updated manuscript

***2. specify temperature for centrifugation ( for example item 1.4.1.3 line 170)***

We have included this edit in our updated manuscript

***3. specify temperature for other steps (for example ethanol precipitation step 1.4.2.1line 177)***

We have included these edits in our updated manuscript

***4. When using thermal cycler for incubations specify whether heat led should be on or off (lines 226,242,265)***

We have included this edit in our updated manuscript

***5. In item 4.3 (line 267) and elsewhere when magnetic beads (such as Agencourt AMPure XP) is used. the buffer concentration/volume affects size selection. It is therefore important to specify the volume/concentration of buffer used.***

We have included this edit in our updated manuscript

1 Yu, M. *et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat Protoc*. **7** (12), 2159-70, doi: 10.1038/nprot.2012.137 (2012).

2 Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet*. **44** (11), 1207-14, doi: 10.1038/ng.2442 (2012).

3 Boyle, P. *et al.* Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol*. **13** (10), R92, doi: 10.1186/gb-2012-13-10-r92 (2012).

4 Chatterjee, A., Rodger, E.J., Stockwell, P.A., Weeks, R.J., Morison, I.M. Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries. *Journal of biomedicine & biotechnology*. **2012**, 741542, doi: 10.1155/2012/741542 (2012).

5 Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc*. **6** (4), 468 - 481, doi: 10.1038/nprot.2010.190 (2011).