# Response to the editor:

Scientific issues:

1. We now require a Data Availability Statement (DAS) with all published papers. Please see the attached PDF for details. All data sources should be mentioned in the DAS, even if the data are available in the manuscript or associated supplementary files.

Response: All Data have been submitted to NCBI GEO.

2. The remaining concerns of referee #2 need to be resolved. We agree with the points in your previous email stating that the analyses should remain in the paper. However, we ask that you explicitly state the potential implications of batch effects on the results, as suggested by the referee, within the results section.

Response: We have revised the main text to explicitly state the potential implications as suggested by referee #2.

3. Please also address the final comment of referee #2 asking you to avoid overstating any claims. We will also carefully check and edit as needed before accepting the final version.

Response: Yes, we have made the necessary edits.

4. Please address all other remaining comments of the referees. These are minor and/or aesthetic.

Response: See point-by-point response in “Response to Referees” section.

General formatting:

The followings are edits to be made to the main text.

5. Article/Technical Report: Our standard word limit is 4,000 words for the Introduction, Results and Discussion. Your current manuscript is 4,431 words, and you will need to cut 431 words. Please contact me if you have any difficulty with this.

Response: We have moved some non-essential text to the Supplementary Notes, and reduced the word count to 4,178 for Introduction, Results, and Discussion combined. Further reduction would affect the essential content of this paper.

6. Online Methods do not have a strict limit, but we suggest 3,000 words as a target. You currently have both an Online Methods section in the main manuscript file and a separate Online Methods in a different file. Please consolidate these and include up to 3,000 words of Online Methods in the main manuscript file. Equation-heavy portions, discussion, or other details that extend significantly beyond 3,000 words should be included in the Supplementary Note.

Response: We have moved the extended methods details to a separate Supplementary Note document and our Online Methods section now have approximately 2,200 words.

7. Your abstract must be fewer than 150 words and should not include citations.

Response: We now have 150 words and no citation in our abstract.

8. The title should provide a clear and compelling summary of the main findings in fewer than 18 words. We suggest the following title: “Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tissue-of-origin mapping from plasma DNA.” Please let me know if you have any concerns.

Response: We have changed the title to the one suggested.

9. Genes must be clearly distinguished from gene products (e.g., “gene Abc encodes a protein kinase,” not “gene Abc is a protein kinase”). For genes, provide database-approved official symbols (for human genes throughout the paper use http://varnomen.hgvs.org/). For the relevant species, use NCBI Gene: http://www.ncbi.nlm.nih.gov/gene. Italicize gene symbols and functionally defined locus symbols; do not use italics for proteins, noncoding gene products and spelled-out gene names.

Response: We have changed to non-italics for the transcription factor names listed in the main text. The remaining are gene names and therefore we used italics for those.

10. For descriptions of variants, use HGVS notation according to the guidelines at http://varnomen.hgvs.org/. Include the accession code for the corresponding reference sequence at first mention of a variant.

Response: All the variants comply with HGVS notation roles.

Figures and Tables:

11. All figures and tables, including Supplementary items, must be cited in the text in numerical order. Please correct the following: Supplementary Table 12 cited out of order (Currently cited after supp table 1 and before supp table 2).

Response: We have removed the out of order citation to Supplementary Table 12 (re-numbered as Supplementary Table 13 in this revision).

12. Figure legends should be concise. Begin with a brief title and then describe what is presented in the figure and detail all relevant statistical information (as described and declared in the supplied checklist), avoiding inappropriate methodological detail.

Response: We have made necessary edits to the Figure legends.

13. Shadings or symbols in graphs must be defined in some fashion. We prefer that you use a key within the image; do not include colored symbols in the legend. Please remove the yellow background shading in Figure 4.

Response: Yellow background shading was removed in Figure 4.

14. Please remove the top/right borders from the plots in Figures 4 and 5.

Response: Borders have been removed from the heat maps in Figures 4 and 5.

15. The pie charts in Figure 5b are not necessary. Please remove them, and place sample size numbers in parentheses next to the appropriate tissue type label on the heatmaps.

Response: Pie charts have been removed from Figure 5b.

16. All relevant figures must have a definition for any error bars.

Response: Bar plots have been converted to box-and-whiskers to show full distributions so we no longer need error bars.

17. Red/green color contrasts can confuse our colorblind readers; please recolor red/green images in figures 1 and 5.

Response: We have removed red/green color contrasts in Figures 1, 5, and S. Figure 2.

18. All bar graphs should be converted to a dot-plot format or to a box-and-whisker format to show data distribution. All box-plot elements (center line, limits, whiskers, points) should be defined. Please convert the bar graphs in Figure 4.

Response: The bar graphs in Figure 4 have been converted to box-and-whiskers.

Statistics and Reproducibility:

19. The Methods must include a statistics section where you describe the statistical tests used. The supplied checklist provides details of which statistics need to be in the Figure legends, and which assumptions and analytical procedures need to be supplied in the Methods. For all statistics (including error bars), provide the EXACT numbers used to calculate the statistics (reporting individual values rather than a range if n varied among experiments) AND define type of replicates (e.g., cell cultures, technical replicates). Please avoid use of the ambiguous term “biological replicates”; instead state what constituted the replicates (e.g., cell cultures, independent experiments, etc.). For all representative results, indicate number of times experiments were repeated, number of images collected, etc. Indicate statistical tests used, whether the test was one- or two-tailed, exact values (NOT for example: <0.05) for both significant and non-significant P values where relevant,F values and degrees of freedom for all ANOVAs, and t-values and degrees of freedom for t-tests.

Response: All the statistic description has been distributed in corresponding sub-analysis sections, respectively. P-value and corresponding test include one- or two tailed has been labelled.

20. The reproducibility checklist must be complete, accurate and up to date.

Response: Yes, we have made the necessary edits.

Supplementary Information:

21. Please rename the supplementary tables to match journal style. For example, “Supplementary Table 1” instead of “Table S1”.

Response: Tables have been renamed in the Excel worksheets.

22. Please use the attached ISI supplementary figures template to prepare your supplementary figures. All other supplementary items (except for Excel files or the Supplementary Data Set [Perl and R codes]) should be included in a single merged PDF file.

Response: We have prepared the integrated supplementary figures using the supplementary figures template and our supplementary notes are now in a PDF file.

23. Any references cited only in the supplementary information should be placed directly within the relevant supplementary files, rather than in the main reference list. For Supplementary Figures, cite any references within the legend text.

Response: We have separated all the references into the appropriate sections.

In addition to addressing these points, please refer to the attached policy and rights checklist, which contains information on how to comply with our legal guidelines for publication and describes the files that you will need to upload prior to final acceptance. You must initial the relevant portions of this checklist, sign it and return it with your final files. I have also attached a formatting tip sheet for you to consult as you prepare the revised manuscript. Careful attention to this guide will ensure that the production process for your paper is more efficient.

# Response to referees:

**Responses to Reviewer #1 comments**

Basically, the authors have addressed most of the comments and criticisms I have raised in my review of the original manuscript. However, there are some issues, which this reviewer likes to mention:

1. Regarding question 6, the addition of Supplementary Table S12a is appreciated. However, sample UCSD.CRC.020 is designated as “benign rectal mucosa”, hence it was not a colorectal carcinoma. UCSD.CRC.029 is a liposarcoma, which does also not correspond to a colorectal carcinoma. UCSD.LC.002 is a metastatic melanoma, thus it is not a lung cancer but a metastasis from a skin tumor. UCSD.LC.015 is “benign fibrous tissue with scar and mild chronic inflammation”. Therefore, some samples appear to be benign, whereas other samples are in fact metastases from other sites. In spite of this, the liposarcoma has clustered together with the colon and the melanoma with the lung (Figure 4); the same applies to the two aforementioned normal samples, which have clustered together with the lung or colon samples. Furthermore, it appears that, according to the values in Table S6, these tissue samples were indeed not identified as “not colorectal cancer” or “not lung cancer”, respectively, by the plasma analyses. Regarding Figure 4d, the authors mention “a small number (N<5) of outliers; do these outliers include the aforementioned samples? These issues should be clarified.

Response: We thank the reviewer for the careful consideration. We noticed the result is consistent that these benign diagnoses (UCSD.CRC.020: benign rectal mucosa and UCSD.LC.015: chronic inflammation) always have lower MHL for the corresponding plasmas compared with average MHL of CCP or LCP. In addition, we did not perform sample clustering in Figure 4, we only did the feature selection to identify MHBs with tissue-specific signals. Finally, the outliers we mentioned in the manuscript are referring to normal plasmas samples only, since we noticed some normal plasmas have higher average MHL compared with the average value of CRC and LC plasmas. However, we don’t have enough information for these normal plasmas, we cannot make further discussion to the outliers.

2. Does Figure 1a really show a “schematic overview of data generation and analysis”? Regarding the prediction model training, where is the part using 29 WGBS and 14 RRBS data sets? Supplementary Figure 11 is more informative than Figure 1a.

Response: The tissue-of-origin analysis used the prediction model generated by the 43 WGBS and RRBS datasets. In lines 308-310, as well as Figure 5b, we described these datasets. Figure 1a focused on the main parts of this manuscript, while Supplementary Figure 11 was supposed to describe the analyses performed on each dataset. We collected these 29 WGBS (8 our own WGBS, 18 Roadmap and 3 WB) and 14 RRBS (tissue samples of ENCODE including replicates for independent experiments). We realized that Table S12c didn’t list the replicates for each tissue sample, and corrected it so that the replicates and only the samples used for prediction model training are shown. We’ve also corrected the Figure 1a to show the correct numbers.

3. Furthermore, the upper part of Figure 1b was probably copied from the UCSC genome browser, and the text “Txn Factor ChIP” was misplaced within the layered H3K27Ac curves. Instead of “Txn Factor Chip” “Transcription Factor ChIP-seq” might be more informative.

Response: We thank the reviewer for noticing and pointing out this mistake in the Figure1b. We have removed the misplaced text and changed the wording for Figure 1b.

4. The color scale of Figure 1c is still not explained.

Response: The color scale is relative scaled densities (z value in the following *fudgeit* function) while the maximum of the value of the scale is the fourth root maximum local density (default of smoothScatter function in R) within the whole Figure and then transfer the density scales to color scales. In order to reduce the complexity, we remove the bar scales and just indicate the density with the color in which red indicate higher density and blue indicates low density.

fudgeit <- function(){

xm <- get('xm', envir = parent.frame(1))

ym <- get('ym', envir = parent.frame(1))

z <- get('dens', envir = parent.frame(1))

colramp <- get('colramp', parent.frame(1))

fields::image.plot(xm,ym,z, col = colramp(256), legend.only = T, add =F)

}

5. In Fig. 1c the “Primary tumor tissue” panel has a black line above the “1.0” at the y-axis and the dotted line does not seem to be at 0.9. It appears that the dotted and the black lines got out of place. In the legend of Figure 1C what does the following mean: “500,000 adjacent CpG loci in MHB regions were randomly sampling and the attenuation of the the r2 with the distance of the CpG loci in different scenario shown different characteristics.” There is no “yellow dot line” (it is rather a red hue)?

Response: We thank the reviewer for the correction. The black line was removed and the dotted lines was adjusted for the position and we replace the yellow with red in the legend. Meanwhile, we replace our previous statement with “the negative correlation between the r2 and the distance of the CpG loci was observed in different scenario” so that it will be more explicit.

6. In Supp. Fig. 2 the color scale is missing. The red region in the right lower corner indicates that a large number of CpGs are in linkage equilibrium (despite the small distance of 150 bp) and this appears to be very different from the plots shown in Fig. 1C. The authors should comment on this. Figure 3b is not referenced in the text.

Response: To be consistent with Question 5’s change, we prefer to not adding the color scale since smoothed scatter plot usually don’t need color scale to indicate the 2D densities distribution.

As for the Supp. Fig. 2, reviewer 2 have brought up concerns of potential batch effect between different groups of samples. We have revised the text to explicitly state this possibility, and provided evidence that this is likely not the case.

Figure 3b has been referenced in the right context.

7. On page 6 lines 255 and 259: the wording that normal plasma contains “residual” or “low tumor contribution” is awkward. As the blood has been obtained from individuals without known cancer, no plasma DNA fragments should be derived from a tumor. The authors should rather call it something like “low/residual plasma fragments with a tumor MHB signature”.

Response: We thank the reviewer for the great suggestion, and have modified the text.

8. Table S6C has probably the wrong heading, as it summarizes lung cancer samples and not colorectal samples. Abbreviations should be spelled out when they are used for the first time. For example, VMR is first used on page 3, line 98, but only spelled out on page 3, line 132 (this also applies to other abbreviations). LAD and LOCK regions should also be spelled out.  
  
Response: We thank the reviewer for the great suggestion. Abbreviations had been checked again and was spelled out before the usage.

**Responses to Reviewer #2 comments**

The authors have adequately addressed my concerns through their revisions, and I am satisfied with the manuscript proceeding. A few remaining points:

1. Line 39: "A number of studies...". You should provide some citations at the end of this sentence, not comprehensive but to the best such small studies.

Response: Yes, we thank the reviewer’s suggestion. Corresponding references have been added.

2. Line 45: "genome-wide" better than "full-genome"

Response: We prefer to “full-genome”, because in many publications assays like Illumina 450k array was mentioned as “genome-wide” assays. In this study we used WGBS data that cover ~26 millions CpG sites, far more denser than 450k CpG sites. “full-genome” is more appropriate.

3. Line 88-92: I continue to be quite concerned that batch effects could cause these differences, despite the additional analyses. Have you taken care that the read length distribution differences between these experiments performed at different sites/studies might not contribute to this? The p-value is kind of meaningless because it's just driven by the large numbers and doesn't provide me any reassurance that this is not batch. My inclination is to suggest just deleting these analyses because I don't think that they add much to the story and are in my view not super believable if the samples were generated/processed/sequenced at different sites, esp. given the subtlety of the differences, e.g. 94.8% vs 91.2%. If you are not going to delete them, at least explicitly remind the reader that there is a risk of batch effects, however careful you were.

Response: We believe that Line 88-92 properly connects to our previous work, and sets a foundation for the rest of the analysis in this manuscript. The decay of methylation LD as a function of CpG spacing has been investigated in our previous study:

<http://genome.cshlp.org/content/20/7/883>

<http://genome.cshlp.org/content/suppl/2010/04/22/gr.104695.109.DC1/GR-104695_SupplementaryInformation.pdf>  (Supp. Figure 4, all data were generated by ourselves in one experiment, hence no batch effect).

Some differences between the two studies include (i) targeted analysis of 2,020 CpG islands versus whole genome analysis; (ii) Paired-end 36bp reads versus paired-end 100bp reads; (iii) cultured cell lines versus primary tissues (plus some cultured cells). Yet in both studies, we consistently observed that methylation LD extends further in stem cells than in somatic cells. Therefore, we didn’t think that the different patterns among three groups of samples shown in Figure 1C is due to technical batch effects.

Importantly, on the cancer side, Catherine Wu’s group published a Cancer Cell paper in 2014 reporting a phenomenon in CLL called “Locally Disordered Methylation”, which was highlighted by a commentary by Charles Swanton & Stephan Beck.

<http://www.sciencedirect.com/science/article/pii/S1535610814004164>

<http://www.sciencedirect.com/science/article/pii/S1535610814004541>

These two papers sent a message to the community that cancer genome is full of discordant methylation (or high stochastic noise), which would mean our methylation haplotyping analysis that focuses on “locally concordant methylation” is not very suitable to cancer samples. A key message that we want to deliver in Figure 1c is, when we took a completely unbiased approach and examined the full human genome, we did notice a detectible gain of “locally discordant methylation”, yet the majority “87.8%” are still concordant.  On the other hand, while the drop of “locally concordant methylation” from 94.8% to 91.2%, and 87.8% appears small, if we look from the other side, focusing on the gain of “locally discordant methylation”, it went up from 5.2% to 8.8% and then to 12.2%, which is quite substantial. We feel that Figure 1c consolidated two seemingly contradicting views and provide a whole picture. It also explained why methylation haplotype analysis allow us to robustly detect tumor and map the tissue-of-origin in plasma.

Therefore, we feel strongly that this is a critical piece of the story and should not be removed. We slightly modified the text to emphasize the points mentioned above.

4. Line 94 - The sentence starting "Gene Ontology" is not very clear, e.g. what is a 'cancer loss linkage region'?

Response: We replace our previous statement with ‘Gene Ontology enrichment analysis to MHB regions whose r2 is decayed compared with normal shown significantly associated with number of cancer related pathway and functions’ so that it will be more explicit.

5. I suggest a careful reading and editing to be more circumspect particularly around self-promotion of the claims. This always hurts more than helps from the perspective of the reader and you'd do be more cautious. One example is "very accurate prediction" at line 363 (what is justifiably considered very accurate in this context depends on a lot of factors, better just to say 'accurate'). It may seem like a minor point (and is) but I think it would help if you go through carefully and revise other such instances.

Response: We thank the reviewer for the great suggestion and we have gone through the manuscript to identify and correct these minor points.