**To the author:** Please provide your responses to the reviewers’ comments in the table below. Highlighted text simply points out when specific requests are made by reviewers that we want to make sure are not missed. Editorial comments may be used to draw attention to specific points that we feel are especially important or that we do not feel are necessary in a revision. They may also be used to note when a comment is related to another comment by the same or another reviewer. However, we ask that you respond to all reviewer comments as thoroughly as possible. The referees’ comments have been reproduced in full and entered into the below tables by the editor, though the order in which they are presented may have changed. Please feel free to modify this table as needed, but please do not edit the referee comments column in any way.

**Reviewer #1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Comment** | **Editorial comment** | **Author Response** | **Modification made in manuscript (give quotes and page numbers)** |
| 1 | Guo et al. conducted a comprehensive and systematic analysis of the human genome to identify regions with highly coordinated methylation. To this end, they extended the linkage disequilibrium concept, which is commonly applied in population genetics. The authors used public data sets and newly generated whole genome bisulfite sequencing (WGBS) data to establish a map of tightly coupled CpG sites, which they named Methylation Haplotype Blocks (MHBs) and made tremendous efforts to identify tissue-specific methylation haplotype blocks. Finally, they generated a metric, termed Methylation Haplotype Load (MHL), for tissue-specific methylation analysis. They then used this metric to estimate the tumor load in the circulation of patients with cancer and to map the tissue-of-origin in these patients.  The idea to use co-methylated sites has recently been proposed -as noted by the authors- by Lehmann-Werman et al. in a landmark PNAS paper. However, the study by Guo et al. is much more comprehensive and provides extensive data with a superior resolution. It is convincingly demonstrated that MHBs represent a distinct class of genomic feature and the tissue specific index (TSI) and the set of tissue-specific MHBs represent novel achievements, which should be of great interest for the community.  Yet there are several issues, which need to be addressed: | **Many of the comments below ask for additional details or clarification. Please be sure to address all comments in full.** | Reviewer #1 identified a few important omissions in the manuscript. We have corrected them in the revision as well as improved the figure legends. We also generated a supplementary figure describing the flow of analyses to aid in understanding the manuscript. Furthermore, in this revision, we performed additional analyses to answer Reviewer #1’s questions regarding the behaviors of MHBs in tumor tissues. | Supplementary Figure 11 |
| 2 | The authors explain the 61 sample sets used to characterize the global pattern and distribution of MHBs (i.e. 51 sets of published WGBS data, H1 human embryonic stem cells, and 10 adult tissues of 1 donor). In Fig. 1c they describe a reduction of perfectly coupled CpG pairs in primary tumor tissue. To which primary tumor tissue are the authors referring? The origin of the tumor tissues is not explained. When checking the references of the papers describing the 51 published WGBS data sets and the embryonic stem cells, no tumor samples could be found. |  | We used a colon primary tissue dataset from Ziller et al. Nature 2013, a colon cancer cell line from Blattler et al. Genome Biology 2014, and a colon primary tissue and three lung cancer cell line datasets from Heyn et al. Genome Biology 2016 for the analysis in Figure 1C.  These references were listed in the Supplementary Table S12d. In this revision, we corrected the use of “primary tumor tissue” to “tumor samples” in the main-text to line 93 in page 3. | Page 3, line 93 |
| 3 | Page 3, lines 123-124: It should made clear that these methylation blocks relate to those established by WGBS, as the previous paragraph dealt with data generated by RRBS or the HM450K array. |  | We added a clarification that these MHBs were established by the 61 WGBS dataset. | Page 4, lines 124 |
| 4 | This reviewer applied the formula provided for MHL on page 9; however, different values were obtained for methylation frequencies of 50% than shown in Figure 2. The parameter P(MHi) should be explained in more detail. | **Please check and clarify this formula and the parameter P(MHi).** | The introduction to the formula was updated in the main text in this revision. We broke down the calculation steps using the following diagram to help the reviewer understand the process of the MHL calculation for the fourth and fifth panel of Figure 2. In addition, we attached the code (R script) for Figure 2 so that the readers could re-create the data.  Key:  SMA: Successive Methylated Allele  TA: Total Allele   |  |  |  |  | | --- | --- | --- | --- | | Allele | SMA | TA | Value | | C | 32 | 64 | 1/2x1 | | CC | 12 | 48 | 1/4x2 | | CCC | 4 | 32 | 1/8x3 | | CCCC | 1 | 16 | 1/16x4 |   \*For Figure 2, the fourth panel  MHL=(1/2x1+1/4x2+1/8x3+1/16x4)/10 = 0.1625   |  |  |  |  | | --- | --- | --- | --- | | Allele | SMA | TA | Value | | C | 32 | 64 | 1/2x1 | | CC | 16 | 48 | 1/3x2 | | CCC | 0 | 32 | 0x3 | | CCCC | 0 | 16 | 0x4 |   \*For Figure 2, the fifth panel  MHL=(0.5x1+1/3x2+0x3+0.5x4)/10 = 0.12 | Page 9, Lines 412 |
| 5 | Page 4, line 162: “65 WGBS data (including 4 additional cancer WGBS sets)”: Again, it is not clear what cancer sets are obtained in the first 61 samples (see above) and which 4 additional cancer sets were added. |  | These additional four primary tumor tissue methylomes from Heyn et al. Genome Biology 2016 were listed in Supplementary Table S12d, including one colon cancer(GSM1279521) and three lung cancer datasets (GSM1279522, GSM1279523, GSM1279524). | Page 4, Lines 165-166 |
| 6 | Plasma samples from 59 patients with cancer were analyzed. What stage were these patients in (were these metastasized patients and are there estimates about tumor burden) and which tumor entities (e.g. how many CRCs and LCs; were the LCs small-cell lung cancer (SCLC) or non-small-cell lung cancer (NSCLC))? |  | We have included some clinical information on the samples in Supplementary Table S12a. The cancer patient samples came from a newly established UCSD biorepository where the numbers/types of samples and clinical information is limited. We do not have TNM classification for all patients, nor quantitative estimation of tumor burden (other than what we estimated in this study based on MHL).  We do have the information that colorectal cancer samples include primary and metastatic colon cancers. For the lung cancer samples, they are subtyped as non-small cell lung cancer, lung adenocarcinoma, and squamous cell carcinoma with or without metastasis. | Supplementary Table S12a |
| 7 | The methylation-haplotype based analysis was started with analyses of 5 LCs and 5 CRCs with matched primary tumors. In order to identify “cancer-specific highly methylated haplotypes (csHMH)” the authors focused on MHBs with low MHL in the blood, but haplotypes present in primary tumor and plasma and found csHMH in all cancer plasma samples (average=36; associated with 183 genes). What was the range? They then extended their analyses to 49 other cancer plasma samples without matched tumor sample using 65 normal plasmas as background; here on average 60 csHMH were found. Again, a range would be needed; were csHMH found in all samples? In fact, the term “csHMH” is somewhat misleading: as detailed further below in the text, these csHMHs were also found in considerable percentages of normal plasma samples (13% and 26% of normal plasmas had low tumor contribution from CRC and LC). Thus, they are not really “cancer specific” but rather “cancer associated”. Later the authors describe additional steps to improve sensitivity and specificity. They could consider using “csHMH” later, e.g. for “group II MHBs”. | **We agree that a more precise term should be used (see highlighted portion). Please address all other comments here.** | We agree with the reviewer’s suggestion, the term “csHMH” would be better replaced with “caHMH” for cancer associated HMH. In the matched primary tumors and plasma methylation haplotype analysis, the range of cancer associated haplotype was from 15 to 97. And in the extended analysis with all the samples, the range of cancer associated haplotype was from 10 to 327. This information has been added to Supplementary Table S7.    These “caHMH” were not found in all cancer plasma samples. The most frequent caHMH occurred in 53% (16/30, chr18:4455202-4455210, DLGAP1) CRC plasmas, while in LC plasma samples it is 62% (18/29, chr4:47034461-47034488, GABRB1). We added **Supplementary Figure 7** to show the distribution of detected “caHMH” in cancer plasma samples (CRC and LC).  13% and 26% of normal plasmas had low tumor contribution from CRC and LC. These values were estimated by MHL in all marker regions, rather than these caHMH. caHMH represent haplotypes, which are not easy for quantitative analysis. In this study, we reported them to make a point that there is tumor-specific signal in the plasma. For the quantitative prediction of tumor load and tissue-of-origin mapping, we focused on using MHL. | Page 6, Lines 231-246  **New Supplementary Fig. 7** |
| 8 | Page 6, lines 242-246: Upon quantification of tumor load in plasma samples they found 72.0% DNA contribution from white blood cells in both cancer and normal plasma samples, and 2.3% from the primary tumor; 3.0% from normal tissue of origin. This does not add up to 100% and contradicts previous papers from the Dennis Lo group, e.g. the paper by Sun et al. (cited by the authors) where the contribution by WBC (neutrophils, T cells, B cells) was well above 80% in the majority of the cases and often in the range of 90%. The 72% appear to be too low and in order to judge whether the contribution of 2.3% tumor DNA to the plasma is reasonable information about the tumor entities, tumor stage, and tumor burden would be needed. |  | In the deconvolution analysis, non-negative decomposition with quadratic programming would assign the weight (estimated proportion) to each reference (tissues). In our model, we had 10 different normal tissues, and each has various levels of heterogeneity. These normal tissues also share some level of similarity, and might have some low fraction of white blood cells. So, the deconvolution not only assigned weight to (1) blood, (2) tumor, and (3) the normal tissue that tumor came from, but also, (4) various low percentages to 9 other normal tissues. In contrast, Sun et al. PNAS 2015 used only one normal tissue (liver) in their deconvolution. Therefore, the total percentage of (1) + (2) + (3) is less than 100%, and (4) accounts for the discrepancies between our results and what were reported by Sun et al. | Page 6, lines 262-264 |
| 9 | They then identified a “small subset of MHBs that have significantly higher levels of MHL in cancer plasma than in normal plasma”. 81 CRC MHBs and 94 LC MHBs, which increased the diagnostic sensitivity and specificity, were found. In the 10 patients with matched plasma samples how many of these markers were present in both plasma and primary tumor? Are they located in regions with frequent copy number gain, which may explain their significant increase? What is the variability of their level, i.e. are they always significantly increased or only in a subset? How many of these MHBs will be detected in an average plasma sample of a patient with cancer? To understand the improvement in distinguishing plasma from cancer patients from normal plasma samples, a Figure like Supp. Fig. 5 should be added (also for later comparisons with Supp. Fig. 6). |  | This reviewer asked several great questions related to two parts of analysis that are directly relevant but has slightly different angles (Lines 266-277, Lines 279-299). In fact, we have evaluated multiple options (many didn’t yield satisfactory results and hence not mentioned in the manuscript), and presented the two promising ones. We also made some decisions on how far we should go along each angle to keep the main text tight and focused.  For example, in the method described in Lines 266-277, the idea is to simply take some normal plasma and cancer plasma samples, look for the regions that show some consistently differences, and use these regions to make predictions on additional samples. It doesn’t require any understanding of cancer biology, and has few assumptions. It can be applied to essentially any disease as long as there are some systematic differences between two groups of samples. However, it would only work when the training set covers enough diversity of all normal plasma and cancer plasma. For methylation analysis on plasma DNA, this requirement might be difficult to satisfy, due to the technical difficulty in sampling enough molecules per locus (limited cfDNA can be extracted from plasma practically, coupled with sample loss during bisulfite conversion). In the situation that tumor DNA fraction is low cfDNA (typically 0.1%-1% as indicated by many mutation sequencing studies), the chance of consistently detecting tumor signatures across many patients on a few markers is low (we covered this point in the Discussions). There has been more than 15 years of research on methylation-based detection in plasma (some early examples summarized by Laird P, Nature Review Genetics, 2003), yet with only one exception (Septin9 based test for colon cancer) most failed due to limited combination of sensitivity/specificity. With the 81 and 94 MHBs, we achieved very favorable sensitivity/specificity compared with previous efforts. However, that is still not sufficient for real clinical use.  In comparison, the approach we took in Lines 279-299 could more promising, since the markers were identified from normal tissues, and cancer biopsies that have much higher fractions of cancer DNA than plasma DNA. Furthermore, information from all the markers were incorporated into a single predictor and hence it’s more tolerant to limited sampling depth. Therefore, in the section of Lines 266-277 we did not expand the analysis further, and moved on to the next part.  We have prepared a new Supplementary Figure 11 to clarify the data use for each analysis. Here are the answers to several questions from the reviewer:  To check the potential enrichment in regions with frequent copy number gain, we downloaded the regions with frequent copy number gains in colon cancer and lung cancer from Pan-cancer project ([Travis](http://www.nature.com/ng/journal/v45/n10/full/ng.2760.html#auth-1) et al. Nature Genetics 2013). A total of 98 colon cancer and lung cancer frequent copy number variations (gain) datasets were collected and they covered 351.2Mb in human genome. Only 17 (8 CRC and 9 LC) of 176 differential MHL regions (17.5K bp) were overlapped with these frequently CNVs (Gain). So there wasn’t an enrichment.  The most frequent MHBs that have hyper-MHL observed in 30% (9 in 30) and 38% (11 in 29) CRC and LC, respectively. The percentages not being higher could be due to multiple reason, such as the heterogeneity of cancers, and limited sampling depth on low-input materials.  In the 10 patients with matched plasma samples, 87.6% (71 of 81, CRC) and 88.3% (83 of 94, LC) of these markers were present in at least one of the plasma and primary tumor pairs. | Page 6, lines 269-270  Supplementary Fig. 11 |
| 10 | Further improvement was achieved by clustering MHBs with high MHL (>0.5) in primary tumors and low MHL (<0.1) in both whole blood and normal tissue. The thus identified “group II MHBs” also showed high MHL in cancer plasma. This may represent a strategy to identify suitable MHBs for these analyses. Do the authors have an explanation why MHL yield an additional gain of signal-to-noise ratio if the tumor fraction decreases below 10%? The authors observed a “significantly higher” average MHL in cancer plasma than in normal plasma. How was the improvement compared to the previous step, i.e. when a “small subset of MHBs that have significantly higher levels of MHL in cancer plasma than in normal plasma” was selected? |  | Methylation haplotype load was designed to favor long continuous methylation CpG fragment (derived from cancer/dead tissues) compared with same methylation level stochastic methylated fragment (plasma background). As the relative abundance of cancer fragment decreases, the signal of MHL does not decrease that much while the standard deviation become smaller and smaller, which is quite different from average methylation level which is more or less stochastically distributed. Therefore, the signal-vs-noise (mean/SD) increased in our sampling analysis. In our opinion, this is the best illustration of the advantage of methylation haplotype analysis.  This part of analysis used more markers, selected based on additional information from normal tissues, whole blood and cancer primary tissues. In contrast, in Lines 266-277, the 81 CRC MHBs and 94 LC MHBs were identified based on differential MHL between only two groups of samples, cancer plasma versus normal plasma. | None |
| 11 | The reduction of perfectly coupled CpG pairs in primary tumors (Fig. 1c) is intriguing. At the same time, there must be genes, which maintain a high MHL in tumors (otherwise, this approach would not work). Is there a certain pattern among genes with loss or maintenance of their methylation patterns (e.g. in terms of involved pathways, function, regulatory networks)? Is loss or maintenance a random process so that csHMHs have to be newly established for each individual patient for a liquid biopsy approach? What is known about the stability of csHMHs if one wants to use them for longitudinal tumor genome monitoring purposes? |  | In Fig. 1c, we randomly sampled 500,000 adjacent CpG loci in MHBs to investigate the relationship between distance and corresponding r2. We found a noticeable reduction of the linkage (r2) in cancer genomes.  To investigate whether there is any specific set of genes involved in the reduction of methylation linkage, we first examined 35,626 linkage regions that were maintained both in normal tissues and cancer tissues, and found no significant Gene Ontology enrichment. In contrast, for the 22,649 regions that had a reduction of methylation linkage only in cancer, many Gene Ontologies were found significant, including different cancer related functions such as [focal adhesion](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0005925&ontoName=GOCellularComponent&species=hg19&ontoUiName=GO%20Cellular%20Component&foreName=user-provided%20data&backName=&sessionName=20161124-public-3.0.0-LGU1Mx) (FDR<5.7x10-3), [cell-substrate adherent junction](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=GO:0005924&ontoName=GOCellularComponent&species=hg19&ontoUiName=GO%20Cellular%20Component&foreName=user-provided%20data&backName=&sessionName=20161124-public-3.0.0-LGU1Mx)(FDR <7.3x10-3), and [Cell Cycle: G1/S Check Point](http://bejerano.stanford.edu/great/public/cgi-bin/showTermDetails.php?termId=BIOCARTA_G1_PATHWAY&ontoName=MSigDBGeneSetsCanonicalPathway&species=hg19&ontoUiName=MSigDB%20Pathway&foreName=user-provided%20data&backName=&sessionName=20161124-public-3.0.0-LGU1Mx) (FDR<0.012). Therefore, the loss of maintenance was not on completely random genomic regions, and hence there were “windows” in the genome that were not affected by the loss of maintenance. Thus, common csHMHs can potentially be identified for different patients. We have attached these enrichment analysis as **Supplementary Table S1b**.  The stability of csHMHs and the feasibility for long-term tumor genome monitoring is an interesting angle for a follow up study. The sample size in this study is too small to provide a definitive answer. | Page 3, Lines 93 to 95  Supplementary Table S1b |
| 12 | The tissue-of-origin part appears to be very nice and represents an advance over previous studies! |  | We greatly appreciate the recognition! | **None** |
| 13 | General remark: the Figure legends are too short and not informative. For example, the percentages in Figure 1c are not explained (they are explained in the text, e.g. on page 3, lines 91-92, but not in the legend), or the legend to Figure 4c is missing. | **Please be sure that all necessary information is included to interpret the figure legend. Figure legends should be no more than 250 words.** | We have updated the figure legends with more details. | **Page 10-14, Lines 495-594** |

**Reviewer #2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Comment** | **Editorial comment** | **Author Response** | **Modification made in manuscript (give quotes and/or page numbers)** |
|  | In this paper, Guo et al. analyze public and new data (including WGBS, RRBS, array) to define ~150,000 blocks of CpG sites that appear to exhibit tight coupling with respect to methylation status (‘Methylation Haplotype Blocks’ or MHBs). They define a metric termed ‘Methylation Haplotype Load’ (MHL) which they use to identify tissue-specific subsets of MHBs. They further apply this framework in the context of methylation analyses of cell-free DNA, i.e. investigating whether MHL facilitates the estimation of tumor load and tissue-of-origin mapping in cancer patients. The results show considerable promise for methylation markers for cell-free DNA analysis of cancer including for tissue-of-origin detection.  Overall, I enjoyed reading the manuscript quite a bit and found it to be a very interesting piece of work. I do think that it would be a more effective presentation if the key take-homes were more crisply articulated for a general audience. The first part of the paper focuses on MHBs – a clearer argument and more forcefully stated case for the value of MHBs over CGIs in how we think about methylation would improve the manuscript. The second part of the paper focuses on cfDNA – clearer articulation of both the promise and limits of methylation in this context could be made in the abstract as well as the paper itself. But I think that those pieces are there and are potentially at the level of Nature Genetics. There are a number of specific points on which I have suggestions or request clarifications. | **Many of the comments below ask for additional details/clarification, but the referee has also made helpful suggestions to improve the clarity of the paper and its main points.** | We thank Reviewer #2 for these positive comments. We have performed all suggested analyses, which resulted in additional supplementary figures and supplementary tables. | None |
| 1 | It seems likely that the dominant source of correlated CpGs are CpG islands (CGIs) which are of course a well-documented phenomenon. So to me where this gets most interesting are both correlated CpG blocks that are not parts of CGIs), as well as places where correlation between nearby CpGs breakdown. | **See next comment** | We performed the analysis to look at the breakdown of the 150k MHBs in different CpG densities as suggested (see below) | **None** |
| 2 | For the former (correlated CpG blocks), it would be nice to see more breakdowns about the ~150K MHBs. What is the full size distribution (histogram)? What is the distribution of CpG densities? (ideally visualized for separately for CGI and non-CGI, as well as overall). Fig 1E lacks nuance beyond CGI vs. non-CGI (which is more of a continuum than is usually let on rather than a dichotomy). If you instead (or additionally) break it down several CpG density bins (rather than CGI vs. non-CGI), how does it look? For example, it looks like non-CGI MHBs are enriched for enhancers – how does this enrichment look if you break MHBs down into a series of CpG density bins (more bins than just CGI vs non-CGI)? | **Please include these data. You can present the histograms and any other visualizations in the supplementary materials.** | We agreed with the reviewer that breaking down MHBs into different levels of CpG densities would tell us how much of the enrichment for different features can be attributed to being parts of CpG islands. Thus, we broke the MHBs into quartiles where the CpGs/bp of each quantile are as follows: (0,0.046), (0.046,0.096), (0.096, 0.155), and (0.155,0.6). We found that while the 1st bin contains only 3% of the total number of CpG island overlapping MHBs, it covers 50% of CpG shelf MHBs, and contributes the most to Fantom enhancers, LAD, and LOCK elements. Notably, the number of Super-enhancer and TAD overlapping MHBs were similarly distributed across the bins, which was a reflection of these features having no known CpG density specificity. | New Supplementary Figure 1 |
| 3 | For the latter (‘breaks’ between blocks, i.e. where nearby CpGs not correlated), Fig 1B is fascinating in showing groups of CpGs with what appears to be a sharp boundary in terms of their correlation structure. Is this the exception or the rule? What features correlate with such boundaries? Have you confirmed this is not a power issue (e.g. if you are lacking reads that traverse the boundary, for example). Additional browser shots (or a browser track) of the haplotype structures would be helpful in this regard. | **Please address the question about power by including the details of any power analysis.** | Figure 1b was an example of a MHB where smaller adjacent MHBs have been merged together to form one large one despite the lack of reads coverage across the ‘breaks’. These breaks or regions with lack of reads coverage do not define MHB boundaries. To show that our MHB boundaries were not defined by lack of coverage, we calculated the total depth of coverage for CpGs which were within a MHB and for CpGs within 100 bp of a MHB. Both sets of CpGs have a median coverage of 1000x. We also prepared a methylation haplotype block structure plot in a quite long genomic region (10Kbp, RASSF1: chr3:50373531-50384063) to show the methylation haplotype linkage block structure in a larger genomic window. | See “Only-for-Reviewer-Figure 1.pdf”, “Only-for-Reviewer-Figure 2.pdf” |
| 4 | Why are only 58.2% of autosomal CpGs covered despite >2,000x coverage? Is that because of mappability issues or the non-covered CpGs being isolated and therefore unable/unlikely to contribute to an MHB? Could be clearer in text. | **We agree that this should be explicitly addressed in the main text.** | We did discuss the reason in one sentence after, but perhaps that wasn’t clear enough.  “The regions not covered by such blocks have low CpG density and hence too few CpG sites within Illumina read pairs for deriving informative haplotypes.”  The key point was that we only considered the CpG sites within “haplotype informative reads”, which required the presence of at least 2 CpG sites within a sequencing read (or paired-end read). Many CpG sites in CpG sparse region were excluded for not contributing to haplotype informative reads. Furthermore, we excluded genomic regions with low mappability. These two “filters” led to only 58.2% of autosomal CpGs being used in our analysis.  We have modified the text. | Page 2, Lines 84-86 |
| 5 | I worry about whether systematic/technical factors (i.e. the numbers of samples falling into stem vs. normal vs. cancer categories; where and how these samples were processed and sequenced) influence the analysis of differences between MHB length and/or correlation strength. Did the authors check that such factors were not contributing? The same question could be asked about Fig S3. Do the cancer and stem cell samples come from the same source/pipeline as at least some of the adult tissues? |  | While the cancer, stem cells, and adult tissues were not from the same laboratories, we performed all the analysis starting from raw fastq files using the same pipeline in order to minimize technical factors. Thus, only the factors influenced by the library construction and the sequencing of each sample by the different laboratories could be contributing.  All of the libraries were generated using standard WGBS protocols which included random shearing of the genomic DNA and shotgun sequencing using the Illumina platform. Thus, variabilities in genomic coverages were completely random. This leaves just the bisulfite conversion efficiencies as a non-random technical factor which we’ve found to vary just slightly among the 65 WGBS libraries (Supplementary Table S12h; Only-for-Reviewer-Figure 3.pdf). Even in Supplementary Figure 3, we see that while not all colon tumor samples were from the same laboratory, they clustered together. Similarly, the adult tissues were from 2 different groups but they also clustered together. The methylation haplotyping analysis and the definition of MHL is quite robust against batch effects. In this revision, we also added two new kidney cancer data sets reported in a separate publication, and observed consistent patterns (new Supplementary Figure 2). | See “Only-for-Reviewer-Figure 3.pdf”; new Supplementary Figure 2 |
| 6 | The statement that “the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs” would benefit from more explanation. |  | Correlated methylation blocks were defined based on the average CpG methylation level for each sample. In contrast, MHBs were defined based on the methylation patterns on single DNA molecules, not the average of all molecules in each sample.  Determination of a linkage equilibrium statistic for a pair of CpGs requires the presence of both a methylated and unmethylated haplotype for both CpGs. This is met for all pairs with correlated methylation. However, not all pairs of correlated CpGs can lead to a high linkage equilibrium statistic. For example, a set of 10 MM, 10 MU, and 10 UU haplotypes (for CpG#1 and CpG#2) are found from three different samples which have an r2 value of 0.250. If sample 1 contributed 5 MM and 5 MU haplotypes, sample 2 contributed 5 MM and 5 MU haplotypes, and sample 3 contributed 10 UU haplotypes, their average methylation levels are 1, 1, and 0 for CpG #1, and 0.5, 0.5, 0 for CpG #2. These CpGs will have perfect correlation but a low r2. | None |
| 7 | I’m confused why MHBs are enriched in enhancers (Fig 1E) but then don’t correlate coherently with enhancers (Fig S2)? |  | In Figure 1D & E we showed that MHBs have significant overlap with enhancers (a static definition with no cell-type specificity), although only about 6% (~9000) of MHBs are overlapping with enhancers. This led us to look further into whether MHBs were also marked by histones which demarcate **active** enhancers (H3K4me1 and H3K27ac) in the specific tissue types, and not too surprisingly, they do not correlate. This is due both to the weak overlap and the differences in enhancer definitions. | None |
| 8 | I think that greater emphasis on the cfDNA results in the overall paper would be warranted (although as per my comments 1-3 above, I think that more could be done to show that MHBs are a more powerful concept than CGIs for investigating methylation-based genomic regulation). However, there are a number of details that would be helpful to add (see below). For example, more details on what stage cancer these patients are at (i.e. are these mostly or all late stage?) should be provided in the main text. | **Additional discussion on the relative utility of MHBs compared to CGIs could be added; however, we are not asking for additional experiments or analyses at this time, except where an analysis is needed to respond to specific requests for clarification (as in comment 10 below).** | Clinical information related to the 59 cancer patients were provided in Supplementary Table S12a. They were not all late stage cancer patients, but included some metastatic ones. | None |
| 9 | The paragraph of text at (lines 240-254) needs some tweaks. For example, not sure what the 95% Cis are referring to – these seem like they should be ranges, not confidence intervals (e.g. I imagine estimated tumor contribution varies by patient). Where is the other 28% assigned? Fig S5 (box plots of % cancer fragment proportion) (and perhaps Fig 6B) should really be moved to main text if possible (this is a really important figure for the field – and illustrates both the potential and the noise). Can the dots be colored by stage? Other places in this paragraph need ranges, the 0.17% and 1% figures. |  | Yes, we assumed that the contribution of the WBC DNA in the plasma would follow the Gaussian distribution, therefore, 95% confidence interval of the contribution were estimated based on the variation among the different samples (cancer plasma and normal plasma, respectively, Supplementary Table S6). We didn’t apply the full range to show variance since it would be strongly influenced by the outlier of clinical samples in which the contribution from WBC is <10%. Such outliers might be caused by different reasons (sample DNA quality, sequencing coverage, unexpected clinical situations, etc.). We appreciated the reviewer comments and we replaced the 95% CI with interquartile range (IQR) to show the estimation to the contribution of WBC and cancer DNAs. As to the other 28% assigned contribution, it was different among different samples, where the major contributions were from cancers, liver, and stomach. | Page 6. Lines 253-258 |
| 10 | In general, the authors appear to take care to validate their classifiers on a test set that is entirely independent of the data on which they trained. However, this could/should be made even more explicit for each analysis, as it’s obviously really important. For example, at lines 256-266, are you measuring diagnostic sensitivity on a set of test samples that is entirely independent of the set that was used to pick those diagnostic MHBs? | **Please include details on validation as requested.** | Yes, we’ve performed cross validation using completely independent test and training data. A description was added in the method section to be more explicit. | Page 10, Lines 444 to 445 |