We thank both referees for their positive comments and useful suggestions for improving our manuscript. In this work, we have generated a catalog of regions with highly coordinated methylation and developed a metric called the methylation haplotype load (MHL) to perform tissue specific methylation analysis. We further demonstrated that due to its increased sensitivity towards long methylated haplotypes, MHL could be used to estimate tumor load and perform tissue of origin mapping. Finally, we applied MHL to perform cancer detection via the cell-free DNA from plasma samples collected from 75 controls individuals and 59 cancer patients. We were pleased to see that the referees found no major issue with our manuscript and they have also raised important questions and suggested a reasonable number of analyses to address those questions.

In particular, both referees wrote positively about our cancer detection methodology via cfDNA and asked that more clinical information for the 59 cancer patient samples be provided. We agree with the referees that this information would be helpful to better understand how our method could be applied in a clinical setting. However, this work was a proof of concept and we had very limited access to samples and clinical information about each sample. Ideally, the variabilities among the patient samples that we observed could be better understood through stratification of the patient samples, and in a larger study with many more samples. Nonetheless, this work have demonstrated that it is feasible to utilize cfDNA to detect cancer in patients with good specificity and sensitivity. Furthermore, we’ve identified potential areas where technical improvements such as higher complexity sequencing libraries from cfDNA samples will help bring us closer to clinical usage.

With regard to inter-sample differences, particularly between progenitors, adult tissues, and primary tumor tissues, the referees were correct to point out potential technical factors such as sample size, sample preparation, sequencing, and data processing. The WGBS used for Figure 1B and Figure S3 were produced by a number of different groups, as follows: the progenitor cells were all from Xie et al. 2013, the adult tissues were from Schultz et al. 2015, Heyn et al 2013, and 10 adult tissue WGBS were by us, and the primary tumor tissues were from Ziller et al. 2013 and Heyn et al. 2016. We have processed all the datasets using the same analysis pipeline to reduce technical factors. All of the samples were sequenced using the Illumina platforms and we’ve performed adaptors and quality trimming to remove low quality base pairs from the reads. We also tested for correlations between principle components in Figure S3 and technical factors such as bisulfite conversion rate and genome coverage and found none to be confounding (Supplementary Figure X2). For Figure 1B, the primary tumor tissues were comprised of 3 colon cancers and 3 lung cancers together and therefore we do not know if the reduction of perfectly coupled CpGs could be extended to other types of primary cancers. To check for whether additional primary tumor samples will follow a similar trend, we added two kidney cancer samples to the analysis and found even further reduction of coupled CpG pairs (Supplementary Figure X3).

In the revision, we have made all of the corrections in the references, figures, and figure legends and generated 3 additional supplementary figures and 1 additional supplementary table. In the following, we address the reviewers’ comments point-by-point.

Reviewer #1:  
Remarks to the Author:  
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.

Reviewer #1 had identified a few important omissions in the manuscript so we have corrected them in the revision as well as improved the figure legends.  Furthermore, in this revision, we performed analyses to answer the questions regarding the behavior of MHBs in tumor tissues. We also showed in a more rigorous simulation analysis how usage of the MHL can lead to a gain of information over conventional average methylation levels when attempting to distinguish cancer tissues (updated Figure 4).

1. 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.

Response:  We used a colon primary tissue sample from Ziller et al., 2013, Nature and primary colon and lung cancers from Heyn et al 2016, Genome Biology for the analysis in Figure 1C.  These references have been listed in the supplementary Table S12. We have corrected the references in the main-text to line 95 in page 3.

2. 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.

Response: We have made the necessary clarification in the main text (line 124).   
  
3. 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.

Response: The introduction to the formula was updated. We show the step by step values in the table below to demonstrate the process of the MHL calculation. (Figure 2, most right panel). In addition, we attached the raw code (R) for the Figure 2 so that readers could make further extensive research.

|  |  |  |  |
| --- | --- | --- | --- |
| 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, most right panel

SMA: Successive Methylated Allele

TA: Total Allele

MHL=(0.5x1+1/3x2+0x3+0.5x4)/10 = 0.12

4. 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.

Response: We’ve added the reference to Heyn et al. 2016, Genome Biology in line 162.

5. 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))?

Response: We have TNM classification for some but not all of the patients. We noticed that colorectal cancer samples are including primary and metastatic colon cancers. For the lung cancer samples, they are belonging to non-small cell lung cancer and include lung adenocarcinoma and Squamous cell carcinoma with or without metastasis. We don’t have other measurement for the estimation of tumor burden. As we known, till now, there is no gold standard for evaluation of tumor burden. Therefore, we did the estimation the tumor burden in our study using non-negative decomposition with quadratic programming based on our MHL matrix which have been shown in the Supplementary Figure 5.

6. 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”.

Response: We agree with the reviewer’s suggestion, the term “csHMH” would be better to be replaced with “caHMH” for cancer associated HMH. In the matched primary tumors and plasma methylation haplotype analysis, the range of the cancer associated haplotype is from 15 to 97. And in the extended analysis with all the samples, the range of the cancer associated haplotype is from 10 to 327.

7. 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.

Response: In the deconvolution, non-negative decomposition with quadratic programming would assign the weight (estimated proportion) to each reference (tissue). In our model, we do have 12 different references/tissues and therefore, the sum of the contribution from blood and normal/cancer tissue is less than 100%. On the other side, we also noticed the paper mentioned by the reviewer which published by Dr. Sun and Dr. [Dennis Lo](http://www.pnas.org/search?author1=Yuk+Ming+Dennis+Lo&sortspec=date&submit=Submit) in 2015 on PNAS. They claimed that ‘One of the key observations from this work is that DNA derived from white blood cells (i.e., neutrophils and lymphocytes) typically contributes more than 70% of the circulating DNA pool, sometimes even to more than 90%’.  In addition, we also notice in Dr. Lo’s PNAS paper, the variation of the deconvolution is also highly variable (Figure 7 in PNAS), which the liver contribution in the blood ranges from about 15% to 70%.  Considering that the diversity of the cancer samples, disease progress and some other confounders, such kinds of difference should be acceptable. We speculated that these high variation is caused by the mixture characteristic of features in the deconvolution analysis, such as for lung tissues, where there are about 40 different type of cells. In order to provide high accurate deconvolution might be helpful to introduce the single-cell methylome reference which is our plans for the next steps.

**8**. 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).

Response: We downloaded the frequent copy number gains of 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, 2013, Nature Genetics 45, 1134–1140).  98 colon cancer and lung cancer frequent copy number variations (gain) was collected and they covered 351.2Mb of the human genome. Only 17 (8 CRC and 9 LC) of 176 identified significantly differential MHL regions in our study (17.5K bp) were overlapped with these frequently CNVs (Gain) and demonstrated CNV gain is not the reason caused hyper-MHL in cancer plasma. The variability of MHL in case are quite high, only few MHL show significant high value in all the cancer samples while majority of the MHLs only with hyper-MHL in subset of the samples which might be caused by complicate reasons, such as heterogeneity of cancers and stochastic release process of cancer DNA to the circulating system. (Supplementary Figure Here, heatmap plot)

9. 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?

Response: 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 concentration of cancer fragment decrease, the signal of MHL didn’t decrease too much while the standard deviation become smaller and smaller which is quite different from average 5mC which is stochastically distributed. Since this analysis (Figure 4C) was terminated at 0.1%, in which the cancer hyper-mhl fragment would be sampled with high probability. We conducted further deep analysis for this simulation analysis (here make it as Figure or supplementary Figure?). We found the distinguish ability for MHL will be disappeared when the cancer DNA proportion ~0.01% while average 5meC based distinguish lost its prediction performance at about  ~0.5%.

**10**. 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?

Response:  In the Fig. 1c, we were randomly sampling 500,000 adjacent CpG loci in MHB regions and the the distance between the CpG site as well as the r2 were recorded to investigate the relationship between distance and corresponding r2.  It is interesting to find the significant decreasing of the linkage (r2) in cancer genomes. Since these regions were sampling from MHB regions randomly, this characteristic should be generalized across the whole genome.

You are quite right. Although the linkage between the adjacent CpG loci have the trending of decreasing, majority of the adjacent CpG loci are still in the high linkage disequilibrium. We build the methylation haplotype block with the same principle based on cancer samples, and we found the MHB number is \*\*\* (should be similar with our initial MHB, right? ) -

In order to reply the question from the reviewer whether some specific region lost/maintain linkage in cancer genome compared with normal or stem cells, we calculate the average r2 to all the MHB regions with the haploinfo data of stem cell, primary normal tissues and cancer tissues and try to identify some cancer associated low-linkage MHB regions. pathway and Gene ontology analysis to these regions shown \*\*\*\*\*\* (we can provide a enrichment figure). Step:  collect and manage mapinfo to 3 group -> merge and concatenate -> resampling to obtain P-value [Provide this result as the supplementary Table]

11. The tissue-of-origin part appears to be very nice and represents an advance over previous studies!  
Response:  Our simulation and real-data analysis showed methylation haplotype load (MHL) performance well for the low level cell-free DNA methylation from the plasma which contains huge number stochastic methylated DNA molecules. MHL could increase the sensitivity and specificity since it could distinguish the continuous methylation CpG loci which is derived from cancer/dead tissue cells from stochastic methylated DNA molecules from the background DNA molecules.

12. 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.  
  
Response:  We have updated the Figure legends with more details.   
  
Reviewer #2:  
Remarks to the Author:  
  
  
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.

We thank Reviewer #2 for these positive comments. In this revision, we performed the analyses to better understand MHBs such as the break down of MHBs into a series of CpG density bins, the enrichment of MHBs at super-enhancers but not at histone marks for enhancers, and to address the various technical issues regarding MHBs identification.

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.

Response: We performed the analysis to look at the breakdown of the 150k MHBs in different CpG densities as suggested below. From the biology respective, these MHB boundaries might be interesting, however, as the biomarker, these regions is not what we are interesting in.

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)?

Response: We agree 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 quartiles are as follows: (0,0.046414], (0.046414,0.096774], (0.096774, 0.15508], and (0.15508,0.6]. We found that while the 1st bin are only 3% of the total number of CpG island overlapping MHBs, it was 50% of CpG shelf MHBs, and contributed the  most to Fantom enhancers, LAD, and LOCK elements. Interestingly, the number of Super-enhancer and TAD overlapping MHBs were similarly distributed across the bins.   
  
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.

Response: We only calculated MHBs for regions with 10X minimum coverage. These genomic blocks were merged if they are within 100 bp of each other.

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.  
Response: Yes, this is mainly due to non-covered CpGs being isolated and the determination cannot be made for it as being a part of an MHB.

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?

Response: 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. These factors could be the bisulfite conversion efficiencies, the genomic coverages, and the read lengths. We tested for correlations between the principle components from Figure S3 to the bisulfite conversion rate and genome coverage and they were not significant (P<XXX, Supplementary Figure X2)  
  
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.

Response: Determination of a linkage equilibrium statistic for a pair of CpGs requires the presence of both a methylated and unmethylated allele for both CpGs. This is met for all pairs with correlated methylation. However, not all pairs of correlated CpGs will obtain a high linkage equilibrium statistic. For example, a set of 10 MM, 10 MU, and 10 UU alleles (for CpG#1 and CpG#2) are found from three different samples which have an LD-R2 value of 0.250. If  sample 1 contributed 5 MM and 5 MU alleles, sample 2 contributed 5 MM and 5 MU alleles, and sample 3 contributed 10 UU alleles, 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 LD-R2.

7. I’m confused why MHBs are enriched in enhancers (Fig 1E) but then don’t correlate coherently with enhancers (Fig S2)?  
TODO: Test whether the enhancer sets used for Fig 1E also have enrichment of the same histones mark for enhancer.

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.

Response: We obtained only limited amount of information about the cancer patients which we included in Supplementary Figure 12a. They are not all late stage cancer patients, and include some metastatic ones.

10. 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.

Response: We have replaced the CIs with IQRs since variabilities among patients may not be normally distributed.

11. 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?

Response: 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.