**Response to Referees**

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.

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.  
  
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.  
  
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.  
  
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.  
  
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))?  
  
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”.  
  
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.  
  
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).  
  
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?  
  
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?  
  
11. The tissue-of-origin part appears to be very nice and represents an advance over previous studies!  
  
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.  
  
  
  
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.

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.  
  
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)?  
  
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.   
  
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.  
  
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?  
  
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.  
  
7. I’m confused why MHBs are enriched in enhancers (Fig 1E) but then don’t correlate coherently with enhancers (Fig S2)?  
  
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.   
  
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.   
  
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?