**Reviewer 2:**

Thank you for the improvements on the manuscript. The introduction portion is greatly improved and most of the questions are reasonably answered by the authors.   
However, the response to question number 6 and 9 required to be clarified further.

***Question 1****: Why defferentially methylated regions (DMRs) are the primary criteria in the biomarker filtering pipeline? If functional significance is not the target of interest and single differentially methylated positions (DMPs) of CpGs are the main targets in the biomarkers, then DMPs could have been identified initially. However, authors funnelled down to single differentially methylated CpGs from the DMRs in the consequent steps.*

**Answer 1**: Thanks for your question. You are right that the functional significance is not the target of interest for our biomarker selection. However, we thought that the single DMPs may be occurred by random and thus reduced its reliability and the possibility for its validation. Therefore, the identification of differentially methylated regions (DMRs) were superior than the DMPs for its robustness due to the fact that several DMPs were included in the same region, indicating the significant hyper/hypo methylation status of these regions. As a result, we first searched for the DMRs but not the DMPs. Moreover, we conducted the targeted bisulfite sequencing methods based on the Next generation sequencing method, which could yield the read length about 200-300 bp, which is short than the length of our DMRs. we could only select 200-300 bp from the DMRs for validation. Therefore, we then selected the most significant DMP in each DMR and validated its adjacent regions for validation. These DMPs would be taken as most optimized/stable biomarkers for the further validation or repeated by other investigators.

***Question 2****: Identification of DMRs is described as taking adjacent six CpGs as methylation block within 1000bp region and used sliding window method to identify DMRs which resulted 105673 DMRs. Why six adjacent CpGs are considered as a cut off to be taken as DMRs.*

**Answer 2**: Thanks for your question. In our study, the definition of the DMR is quite important for identifying the potential ESCC biomarkers. Therefore, many different cutoffs have been applied for DMR definition. We have tried the number from 2 to 12 adjacent CpG sites as the cutoff and plotted the correlation between the cutoff and the number of the DMRs as following. We found that 6 adjacent CpG sites seems as the turning point for the correlation, and the number of DMRs seems to decrease much slower after the 6. As a result, we thought that the choose of six adjacent CpG sites as the cutoff may be adequate for identifying the robust DMRs and will not leave out the potential promising DMRs as the candidate biomarkers. Finally, we take the six adjacent CpG sites as the cutoff to be taken as DMRs.



***Question 3****: If six adjacent CpGs (within 1000bp region) are considered as minimum cut off for DMR criteria and sliding window method results 105,673 DMRs. Then it seems 6 (or more) CpGs multiplied by 105673 will constitute at least 634,038 CpGs, which is not possible as the data only constitutes around 450,000 CpGs. Please clarify it further for better understanding.*

*Overlap*

**Answer 3**: Thanks for your question. We conducted the DMR identification procedures with the sliding window method, which we first sort all the CpG sites according to their genomic locus, and take the 1-6 sites in chromosome 1 as the first DMR, then the 2-7 sites in chromosome 1 as the second DMR. As a result, we could get around 450,000 DMRs. However, some of the DMRs are in a long range (> 1000 bp), and may not share a similar DNA methylation pattern. As a result, we finally obtained 105,673 DMRs for further analysis. To further clarify the method, we have added the detailed description of the sliding window methods as the supplementary Methods as following in Page xx, Line xx-xx:

We then applied the sliding window methods according to the genomic position of each CpG site, starting from the first 1-6 CpG sites located at the chromosome 1 as the first candidate methylation region, and then the 2-7 CpG sites in chromosome 1 until the last six CpG sites in chromosome 22.  
  
***Question 4****: Probably the conversion of DMRs into DMPs (1355) and then funnelling down process to 275 CpGs and then 175 CpGs and ultimately 5 CpGs sites could be elaborated with all the selection criteria into a table or figure (may be in figure 1) for clear and better understanding.*

**Answer 4**: Thanks very much for your suggestion. In the revised manuscript, we have added the details of the selection criteria into the Figure 1.

***Question 5****: One of the sentence say ''It is of great importance that the methylation rate of the candidate biomarker should be very low in the normal tissues as well as in the peripheral blood so that it can be used for non-invasive cancer diagnosis in the future''. Please explain the meaning of low methylation rate in the normal tissues.*

**Answer 5**: Thanks for your question. We feel sorry that we did not say it clear that the ‘normal tissues’ in this sentence represent the normal esophageal squamous cells. As we want to extract the differentially methylated biomarkers for ESCC, we would like the biomarkers to be hyper-methylated in ESCC tumor tissues and the peripheral blood in the ESCC patients, while keeping the low methylation rate in esophageal squamous cells as well as the peripheral blood in healthy samples. To clarify, we have rewrite the sentence in the revised manuscript as following in Page xx, Line xx-xx:

*It is of great importance that the methylation rate of the candidate biomarker should be very low in the normal esophageal squamous cells as well as in the peripheral blood so that it can be used for non-invasive cancer diagnosis in the future.*

***Question 6****: Nothing is mentioned about the hyper or hypo methylation status of the CpGs in the selected CpGs.*

**Answer 6**: Thanks for your suggestion. The aim of our study is to find out the candidate biomarkers for ESCC diagnosis with ESCC tissues and peripheral blood in the future. Therefore, we only want the hyper-methylated CpG sites, which could further be applied for liquid biopsy due to the complexity of the cell composition in the peripheral blood. In addition, in the revised manuscript, we have added the description of hyper/hypo methylation status for the selected CpGs in the abstract and the conclusions.

***Question 7****: Why the CpG sites which are located far from CpG islands were filtered out?*

**Answer 7**: Thanks for your question. Nearly more than 50% of the CpG dinucleotides are located in the CpG island in the human genome. The CpG island have been identified as the important regulators for gene expression and biological functions. As a result, we thought that the hyper-methylation status of the CpG sites located at the CpG islands may be more reliable due to its implications for biological functions. As for the CpG sites far from the CpG islands, their biological meanings are not clear and thus reduced its credibility for becoming the candidate biomarkers. Moreover, the density of the CpG sites are usually higher in the CpG island than the non CpG island regions, for which we could sequence more CpG sites in a candidate DMR if it is located at the CpG island region. What’s more, the DNA methylation status of the CpGs in CpG island are more homogeneous and more consistent with adjacent CpGs and therefore to be good biomarker candidate regions. Finally, the increased number of the sequenced CpG sites could better reveal the actual methylation status in a region. As a result, we filtered out the CpG sites which located far from the CpG islands.

***Question 8****: Why the CpG sites whose corresponding genes have been studied in ESCC carcinogenesis were removed. Though these CpGs were earlier studied but this may not rule out their efficiency as better biomarker.*

**Answer 8**: Thanks for your question. You are right that these CpG sites may be efficient as the biomarker. However, the aim of our study is to find out novel candidate biomarkers for ESCC, and the validation of the previous biomarkers was not in our scope. Therefore, we only selected the CpG sites whose corresponding genes have not been studied in ESCC carcinogenesis. In the next step, we would optimize the technique so that we can allow more regions to be sequencing at the same time and then we can evaluate more regions, including previous reported biomarkers.

***Question 9****: It is also not mentioned how and which form of data (level 1, 2 or 3) from TCGA is downloaded for analysis.*

**Answer 9**: Thanks for your question. In our study, we downloaded the level 3 methylation data from TCGA for analysis. To clarify this, we have added the sentence in Page xx, Line xx-xx in the revised manuscript as following:

Public high-throughput DNA methylation microarray datasets were searched and the comprehensive methylation dataset of esophageal cancer from the TCGA project was the first obtained. There were 84 ESCC and 3 normal tissues in this level 3 dataset. In addition, we also found that there are 78 EAC and 13 adjacent normal tissues in the TCGA dataset (level 3).  
  
***Question 10****: Overall, there are standard up to date methods available for 450k data analysis and identification of DMPs and DMRs which consist of prepossessing and normalization of the data before the regression /differential methylation analysis. These methods also remove SNPs and sex chromosomes during the prepossessing steps which will avoid any biases during the analysis. Authors have removed CpG sites with SNPs after funnelling down to 175 CpGs which is not recommended. Usually the biases due to sex associated CpGs and CpG sites with SNPs could interfere with the differential methylation analysis, therefore, it is advisable to remove them in the prepossessing step.*

**Answer 10**: Thanks for your advice. You are right that several methods have been developed for 450k data analysis and remove the CpG sites with SNPs in primers and located at sex chromosomes first in the prerocessing steps, such as CHAMP and RnBeads. In our pipeline, we first removed the CpG sites located at the sex chromosomes due to the fact that the methylation patterns of the sex chromosomes would be largely affected by the gender and would not be suitable for being as candidate biomarkers. However, we still remained the CpG sites with SNPs in their primers for the reason that we want to extract the methylation profiles of the DMRs but not the single DMP. According to 450K array, there are only ~10% probes which might be influence by SNPs. Therefore, one or two CpG sites with SNPs in their probes would not affect the methylation profiles of DMRs largely. Moreover, our validation stage is based on sequencing technique, genetics variance within the regions will give us great opportunity to investigate the genetic structure or confounder adjustment in the future. Therefore, we didn’t remove the SNP related probes. Our validation stage shown our biomarker selection works well in some way.

***Question 11****: Moreover, in a conventional approach DMPs are initially identified and then DMRs. DMPs are usually significant for biomarker development whereas DMRs mainly shed light on functional relevance of the differentially methylated regions. Here the authors identified the DMRs first and then converted them into DMPs with certain selection criteria, the reason for looking at DMRs for identifying DMPs (differentially methylated single CpG sites) is not justified.*

**Answer 11**: Thanks for your question. You are right that the conventional approach DMPs are initially identified and then the DMRs. However, in our study, we want to first extract the robust DMRs, which the DMPs in these DMRs would be more robust and reliable. However, due to the limitations of the next generation sequencing, we could not sequence all the length of the DMRs and we then selected the most significant DMP in DMR and obtained the targeted regions covering the candidate CpG sites and with the lengths about 200-300 bp. Moreover, the conventional method which identified the DMPs first and then converted DMPs into DMRs, will also face the problem that many DMRs are too long to be sequenced with NGS, and will also need to be shorten to fit the technical limitations of NGS. As a result, we thought that our identification method is robust and reliable for biomarker selection. What’s more, in the future, array based technique might be useful in cancer diagnosis and in these technique stable/consistent probes/CpGs sites will be quite useful.   
  
***Question 12****: The actual tumor percentage is difficult to speculate but there are methods which can assure that the sample processed for DNA/other molecular analysis is extracted from a tissue containing majority of tumor content. Usually the established method could be laser microdissection however due to limitations sometimes pathological reading (more than 50 to 70% tumor content) of the tumor specimen can be considered. Here the objective is to have an idea of the probable tumor content of the processed tumor specimen.  Otherwise, a tumor positive tissue may contain 5% or 95% of tumorous cells which may not show the same molecular feature and it is not reasonable considering them in the same group of tumors.*

**Answer 12**:

Thanks for your question. Theoretically, the quantitative analysis you mentioned is the most accurate way to make the data analysis for biomarker or omic data analysis. However, this concept is only formed recently. Even for the US TCGA project, the cancer purity have not provided. We notice that Dr. Atul J. Butte from UCSF, Dr. Hao Wu from Emory University and Dr. [Xiaole Shirley Liu](https://www.ncbi.nlm.nih.gov/pubmed/?term=Liu%20XS%5BAuthor%5D&cauthor=true&cauthor_uid=25103624) from Harvard have proposed some bioinformatics tools to estimate the cancer purity based on array or sequencing dataset, but it is hard for tissue samples, especially without pre-genome-wide DNA methylation information. Another question is that even we estimated cancer purity for the solid tumor samples, it is also hard to guarantee that the DNA applied in the methylation sequencing is same as the samples we applied for tumor purity estimation. We admit that the idea your mentioned is the future method we need be considered. However, we have proof that with our pipeline some significant biomarkers could be identified successfully.