Circulating cell-free DNA based low-pass genome-wide bisulfite sequencing aids non-invasive surveillance to hepatocellular carcinoma

Shicheng Guo[1, #], Haikun Zhang[2,#], Peiling Dong[2, #], Jiakang Wang[3], Ramsey Cheung[4], Augusto Villanueva[5], Steven J. Schrodi[1,6,*], Dake Zhang[2,*], Changqing Zeng[2,*]

[1]Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, USA
[2]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China
[3]Biology Department, Stonybrook University, Stonybrook, NY, USA,
[4]Department of Gastroenterology and Hepatology, VA Palo Alto Health Care System and Stanford University, Palo Alto, CA, USA
[5]Liver Cancer Research Program, Division of Liver Diseases, Tisch Cancer Institute, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6]Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, USA

Abstract

Circulating cell-free DNA (cfDNA) methylation has been demonstrated to be a promising approach for non-invasive cancer diagnosis. However, the low-level of cfDNA and high cost of whole genome bisulfite sequencing (WGBS) significantly hinders the clinical implementation of a methylation-based cfDNA early detection biomarker. Here we proposed a novel method that we utilized long-region methylation ($\mathrm{Methyl}_{\mathrm{LRM}}$) in low-pass WGBS data (~5 million reads) generated from cfDNA to detect methylation changes. We applied the method to investigate dynamic methylation changes in cfDNA from blood samples of patients with hepatitis, cirrhosis, early and advanced hepatocellular carcinoma (HCC). We found a significant enrichment of differential methylation loci in intergenic and repeat regions, especially in HBV integration sites. Moreover, methylation profiles nearby HBV integration sites ($\mathrm{Methy}_{\mathrm{HBV}}$) were found to enhance the prediction performance. Multiple machine learning models based on $\mathrm{Methyl}_{\mathrm{LRM}}$, $\mathrm{Methy}_{\mathrm{HBV}}$, cfDNA fragment size ($\mathrm{cfDNA}_{\mathrm{size}}$) with five-fold cross-validation demonstrated low-pass cfDNA methylation data provided powerful discriminating ability. The results demonstrate that low-pass cfDNA WGBS could be used as a low-cost approach for early HCC detection in the context of surveillance programs.

Introduction

Circulating cell-free DNA has been a useful technique for non-invasive cancer diagnosis. DNA methylation in circulating cell-free DNA has exhibited particularly promising signals for cancer diagnosis and tissue-of-origin mapping. It is now well-recognized that genome-wide DNA hypo-methylation is a hallmark feature of the human cancer genome and therefore may be usefully applied to cell-free DNA-based cancer diagnosis. However, the amount of circulating cell-free DNA is typically too limited for interrogation with conventional high-depth/coverage genome-wide bisulfite sequencing (WGBS). In this original manuscript, we have proposed a novel strategy to apply low-pass WGBS to monitor DNA methylation levels in cell-free DNA fragments. We have developed a new measurement approach for long-region hypo-methylation which shows utility as a biomarker for cancer surveillance in liver diseases ranging from hepatitis, cirrhosis, early stage HCC and advanced HCC. Our study shows that low-pass WGBS provides a stable and powerful diagnostic tool for HCC. Furthermore, our approach enables an evaluation of the efficacy of surgical intervention for HCC. Interestingly, we present evidence of over-representation of differentially methylated CpGs in HBV integration regions based on our low-pass WGBS approach, providing additional insights into the mechanisms of HCC molecular pathophysiology and may aid HCC diagnosis and clinical decisions. Implementation of this approach is favored by the low cost compared to conventional techniques. Using machine learning, we show that HBV integration-based DNA methylation in cell-free DNA ($\mathrm{Methyl}_{\mathrm{HBV}}$) exhibited excellent predictive performance in distinguishing HCC from other liver diseases. Finally, using the same data, we introduced cell-free DNA fragment size distribution effects into our predictive model yielding a powerful HCC discriminating ability.

Data and Method

Sample: 3 normal individual, 17 hepatitis, 17 cirrhosis, 3 early HCC, 5 late HCC and 9 plasma samples after tumor is surgically removed were collected in this study.

Cell-free circulating DNA methylation: low-pass GWBS data were generated for each plasma cell-free DNA samples. Each samples have approximately 5 million 100bp pair-end reads.

Long range methylation (LRM): HCC genome was divided into 500-Kb, 1-Mb, 1.5Mb, 2-Mb and 2.5-Mb segments. Average methylation were calculated. hyper-LRMs and hypo-LRMs were defined by 3 SD beta-value.

cfDNA fragment size determination and distribution: Unique reads with well alignments to human genome were applied for cfDNA fragment size evaluation. The end positions and start positions were extracted to calculate the median of the cfDNA size for further prediction analysis.

Prediction and machine learning Features including AFP, $\mathrm{Methyl}_{\mathrm{LRM}}$, $\mathrm{Methyl}_{\mathrm{HBV}}$, $\mathrm{cfDNA}_{\mathrm{size}}$ were used in multiple machine learning algorithms, such as random forest (RF) and Neural Network (NN), for cancer vs non-cancer binary classification. Five-fold cross-validation were applied to avoid over-fitting. We applied logistic regression to compare the performance of different models with different features.

Result

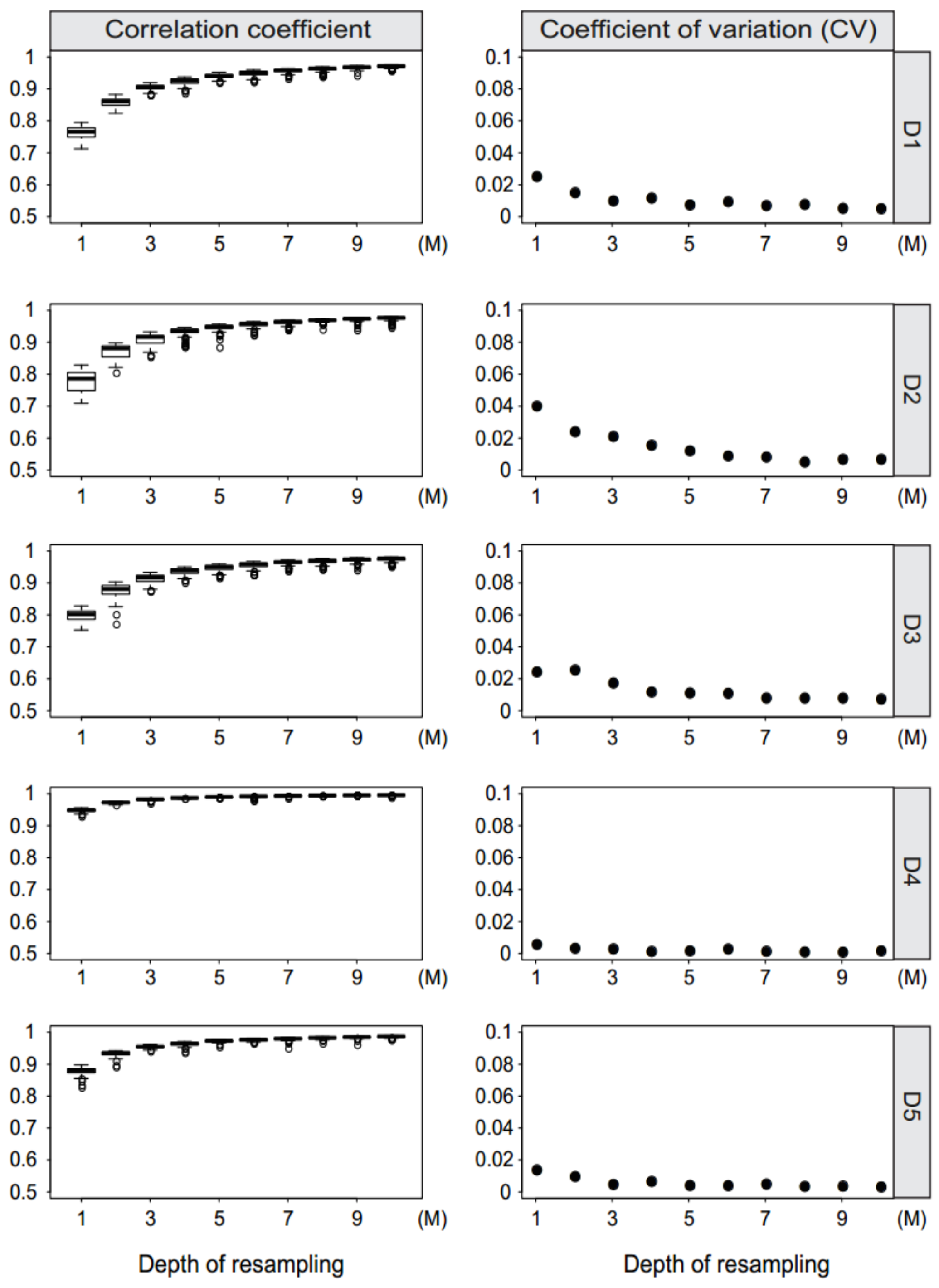

Fig. 1. The efficiency of resampling sequencing reads for low pass WGBS. Left of the figure showed the correlation coefficient between re-sampling low pass WGBS and total sequencing reads 100 times from 1M to 10M. Right panels of the figure show the coefficient of variation (CV) for 100 correlation coefficients between resampling low pass WGBS and total sequencing reads from 1M to 10M.

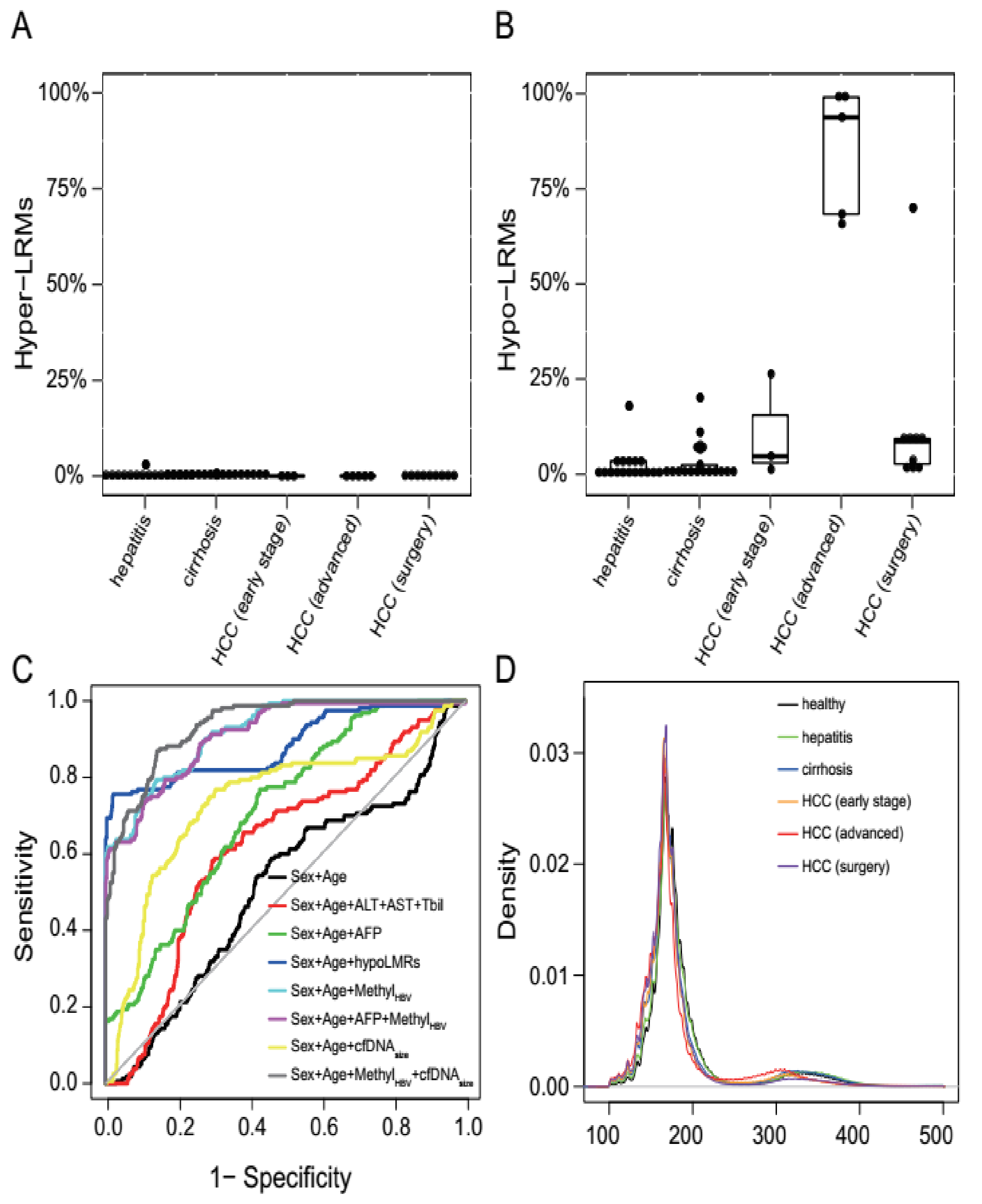

Fig. 2. Percentage of long range methylation (LRM) showing hyper- or hypo-methylation in all the individuals. (A) The percentage of hyper-methylated long range regions (2-Mb) in hepatitis, cirrhosis and HCC patients. (B) The percentage of hypo-methylated long range regions in hepatitis, cirrhosis and HCC patients. (C) Receiver operating characteristics (ROC) curve based on five-fold cross-validation for HCC patient detection by different indicators in discriminating HCC from individuals without HCC. Here, we applied five-fold cross-validation based logistic regression to compare the performance of different models with feature combination. We found $\mathrm{Methyl}_{\mathrm{HBV}}$ and cfDNAsize based prediction model have best performance with AUC=0.94, 95%CI: 0.92-0.96 (D) The distribution of cfDNA fragment size in healthy, hepatitis, cirrhosis, early stage HCC advanced HCC and HCC after surgery. We demonstrate HCC showed short cfDNA median size.

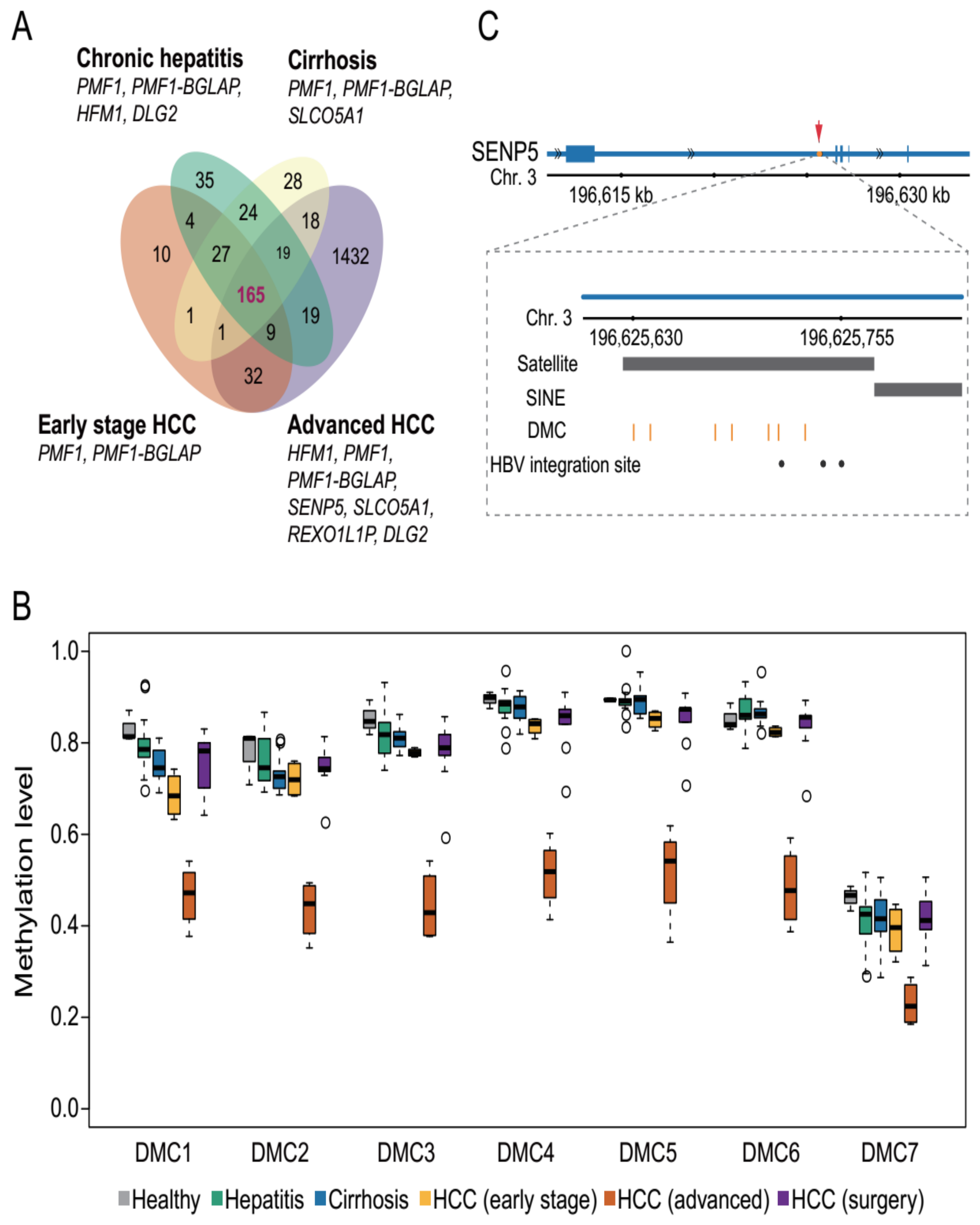

Fig. 3. Differentially methylated CpGs (DMCs) identified by low-pass cell-free WGBS. (A) Left Venn diagram showing the overlap of DMCs generated by 1 hypo-methylated chronic hepatitis patients, 2 hypo-methylated cirrhosis patient, 1 hypo-methylated early stage HCC patients and 5 advanced HCC patients compared to healthy individuals. Genes represent the genes annotated with DMCs in each comparison. (B) Boxplot displays the methylation level of 7 DMCs of SENP5 in all the individuals. (D) The locus of 7 DMCs and 3 reported HBV integration sites in intron 2 of SENP5. The black dots represent the HBV integration sites and the orange vertical lines represent the 7 DMCs. The black bar labels represent the repeat marker locus in this region.

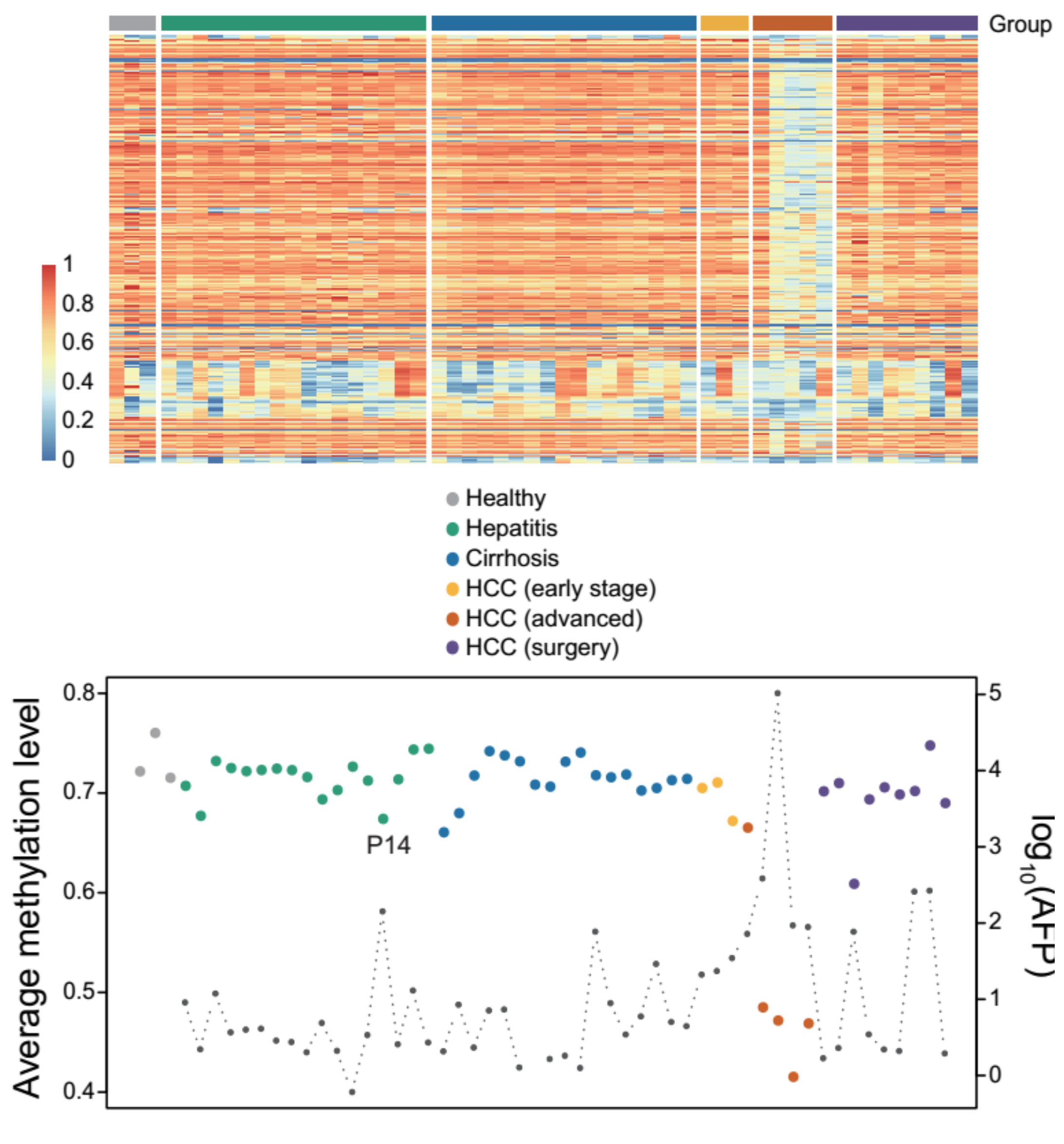

Fig. 4. Overrepresentation of DMCs and CpGs surrounding HBV integration sites. (upper) The heatmap displays the methylation level of the CpGs located within 100bp of the HBV integration sites in all the samples. We found HBV integration sites regions in different samples showed distinct patterns. (bottom) The average methylation level of all the CpGs located within 100 bp of the HBV integration sites in all the samples. The black dot represents for AFP level for the corresponding individual. As we can observed average methylation level start to be decreased in the early HCC and showed significant lower values in late HCC. It is interesting cancer remove surgery could recover the methylation levels

In summary, we demonstrate that $\mathrm{Methyl}_{\mathrm{LRM}}$, $\mathrm{Methyl}_{\mathrm{HBV}}$ and $\mathrm{cfDNA}_{\mathrm{size}}$ could serve as effective features in predictive models to detect HCC. We also demonstrated LRM reflects genome-wide demethylation changes from non-tumoral tissues to HCC and could be used as a low-cost approach detect minimal tumoral residual disease after surgical resection. Our study provided a novel low-cost HCC cancer diagnosis strategy in which HBV integration, DNA methylation and cfDNA fragment size were employed