LiBis: An ultrasensitive alignment method for low-input bisulfite sequencing

Yue Yin[1#], Jia Li[1#], Jin Li[1], Minjung Lee[1], Sibo Zhao[2,3], Linlang Guo[4], Jianfang Li[1,5], Mutian Zhang[1], Yun Huang[1], Xiao-Nan Li[2,6], Deqiang Sun[1,7,8]

[1]Center for Epigenetics and Disease Prevention, Institute of Biosciences and Technology, Texas A&M University, Houston, TX 77030.

[2]Texas Children's Hospital, Baylor College of Medicine, Houston, TX 77030.

[3]Department of Hematology and Oncology, Cook Children's Medical Center, Fort Worth, TX 76104.

[4]Department of Pathology, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, 510700.

[5]The Fifth Affiliated Hospital of Guangzhou Medical University, Guangzhou 510700.

[6] Department of Pediatrics, Ann & Robert H. Lurie Children's Hospital of Chicago, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611.

[7]Department of Epigenomic Products, Genomics Technology Corporation, Houston, TX 77030.

[8]CPRIT scholar for cancer research

#Equal contribution

Correspondence: dsun@tamu.edu

Keywords

Methylation, bisulfite sequencing, WGBS, liquid biopsy, cfDNA, cerebrospinal fluid, virus insertion site

Abstract

The methylation profile of cell-free DNA (cfDNA) in liquid biopsies has been utilized to diagnose early-stage disease and predict therapy response. However, in typical clinical settings, only very small amounts of cfDNA can be purified. Whole-genome bisulfite sequencing (WGBS) is the gold standard to measure DNA methylation; however, WGBS using small amounts of fragmented DNA introduces a critical challenge for data analysis, namely a low mapping ratio. This, in turn, generates low sequencing depth at individual CpG sites and low coverage for CpGs sites genome wide. The lack of informative CpGs has become a bottleneck for the clinical application of cfDNA-based WGBS assays. Hence, we developed LiBis, a novel method for low-input WGBS data alignment. By dynamically clipping initially unmapped reads and remapping clipped fragments, we judiciously rescued those reads and uniquely aligned them to the genome. By substantially increasing the mapping ratio by up to 88%, LiBis improves the number of informative CpGs and the precision to quantify the methylation status of individual CpG sites. The high sensitivity and cost effectiveness afforded by LiBis for low-input samples will allow the discovery of genetic and epigenetic features suitable for downstream analysis and biomarker identification using liquid biopsy.

Introduction

Liquid biopsy has rapidly emerged as an alternative to tumor biopsy in early-stage cancer diagnosis and therapy response monitoring because it is minimally invasive and enables repeat sampling. Of particular utility for liquid biopsy, Circulating cell-free DNA (cfDNA) harboring specific biomarkers, such as single-nucleotide mutations and methylation aberrations, has been demonstrated to have strong potential for cancer diagnosis or screening(1). The cobas *EGFR* Mutation Test has been approved by the FDA to use cfDNA from patient plasma to identify patients with metastatic non-small cell lung cancer who are eligible for treatment with Tarceva (erlotinib)(2). Besides genomic features, epigenomic modifications are also utilized in cancer screening. Epi proColon screens for colorectal cancer by evaluating the methylation status of the *SEPT9* promoter using patient plasma(3). In addition to PCR-based technologies, deep sequencing of cfDNA is widely utilized to search for new candidate biomarkers. However, two challenges exist for sequencing cfDNA for biomarkers. One challenge is that the amount of cfDNA collected from the plasma of healthy individuals or patients usually ranges from a few nanograms to a few dozen nanograms, and for late-stage cancer patients the amount is usually less than several hundred nanograms. Another challenge is that cfDNA is highly fragmented, mostly around 120 bp, and gets substantially smaller after library preparation(4); these small fragments are not suitable for the most economical sequencing method, which is the high-throughput 150 bp paired-end sequencing mode on the Illumina NovaSeq platform.

Sequencing library preparation methods have been developed for low-input DNA samples to improve the sensitivity of whole-genome sequencing. For example, cancer personalized profiling by deep sequencing can quantify circulating tumor DNA(ctDNA) using as low as 4 ng DNA as input by hybridizing specific regions of the DNA to specific selectors designed for different cancer types(5). Another method, traditional whole-genome bisulfite sequencing (WGBS), requires a large amount of input DNA because the bisulfite conversion produces additional digestion of both the DNA fragments and the primers. Pre-bisulfite library preparation methods generate the DNA library before applying bisulfite conversion,

which transforms unmethylated cytosine to uracil which is further transformed to thymine during PCR amplification(6-8). Pre-bisulfite libraries require a large amount of starting DNA due to unequal digestion of the various DNA fragments and introduces bias to the methylome analysis(9,10). To efficiently solve this issue, post-bisulfite library methods were recently developed, which apply bisulfite conversion prior to making the libraries(11). In such processes, random primers are used to add adapters to the bisulfite-transformed single-strand DNA. Methods enabling the use of less starting DNA (<100 ng) with less sequencing bias enable researchers to decipher the methylomes of otherwise challenging substrates, such as cfDNA and single cells(12,13). Single-cell WGBS (scWGBS) methods were developed to reveal epigenetic differences at the single-cell level with low amounts of starting DNA. However, low-input DNA bisulfite sequencing methods using dNTP mixtures suffer from the formation of synthetic sequences, and this issue is even worse for post-bisulfite libraries made from small fragments of cfDNA. Both facts lead to low mapping ratios, which limits the data efficiency for thorough exploration(14).

Several mappers such as BSMAP, Bismark, and BSSeeker have been developed to align raw bisulfite sequencing data, and these have been widely used(15-17). Based on BS-Seeker2, which supports soft-clipping in its original procedure, scBS-map was developed to improve the data efficiency of single-cell methylome analysis using a local alignment strategy(18). However, we still observed a low mapping ratio, low sequencing depth, low number of informative CpGs, and insufficient accuracy for low-input bisulfite sequencing, especially for cfDNA. Such bottlenecks need to be removed so that downstream analysis can achieve high mapping ratios for low-input bisulfite sequencing data, high accuracy for methylome profiling, and time-efficient analysis of clinical cfDNA WGBS data and thousands of single-cell methylomes.

Here, we developed LiBis to analyze low-input bisulfite sequencing data. We first present our methodology and then test its performance using one simulation dataset and three real datasets. LiBis dynamically realigns unmapped reads clipped in a sliding window and removes potential effects of sequencing bias and random priming. In the simulation study, we were able to define a ground truth, and

in the first real dataset we were fortunate to be able to assume a ground truth. LiBis can improve the number of usable CpGs and the methylation ratio accuracy with high time efficiency. By applying LiBis, we achieved better cost efficiency using both a public dataset and in-house datasets. The number of informative CpGs increased significantly compared to direct alignment after trimming. The precision of bisulfite sequencing was also improved by LiBis for all samples. Furthermore, LiBis identified virus insertion sites and virus integration mode in a cervical cancer WGBS dataset, which indicates that bisulfite sequencing data can be used to reveal both genetic and epigenetic patterns. LiBis supports a one-command solution for quality control, trimming, mapping, and methylation calling with a short run time, making it a comprehensive solution to support large-scale single-cell or cfDNA bisulfite sequencing applications.

Materials and Methods

Patient sample collection

Signed informed consent was obtained from all patients or their legal guardians prior to sample acquisition in accordance with an institutional review board-approved protocol. CSF samples were obtained from three brain tumor patients at Texas Children's Hospital at the time of clinically indicated lumbar puncture. CSF was processed using a standardized protocol and was then divided into aliquots and stored immediately at -80°C. The cervical tumor tissue was obtained from the Southern Medical University as an FFPE sample.

Tumor DNA and CSF cfDNA purification

Total DNA containing cell-free DNA (cfDNA) was isolated from 200–400 µL CSF using a QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer's instructions. For tumor tissue, DNA was isolated using an AllPrep DNA/RNA Mini Kit according to the manufacturer's protocol. The isolated DNA concentration was measured using a Qubit 4 Fluorometer with the Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific).

WGBS library preparation

WGBS analysis was used to access the genome-wide DNA methylation profile. The cfDNA WGBS libraries were generated using a Pico Methyl-Seq Library Prep Kit (Zymo Research). Briefly, ctDNA was mixed with 0.1% unmethylated λ-bacteriophage DNA (w/w) (NEB), followed by sodium bisulfite conversion. The bisulfite-converted DNA was then annealed with random primers for initial amplification, followed by adaptor ligation and final amplification with Illumina TruSeq indices. Constructed libraries were run on a 2% agarose gel to assess size distribution, and the library concentration was measured using

a Qubit 4 Fluorometer with a Qubit dsDNA High Sensitivity Assay Kit. Normalized libraries were pooled at an equimolar ratio and sequenced on a NextSeq 500 or a NovaSeq 6000 (Illumina).

LiBis implementation

LiBis is available on GitHub (https://github.com/Dangertrip/LiBis). LiBis was developed in Python 3.6 and integrated with published software for infrastructure functionalities. In the clipping mode, reads discarded in first-round mapping are cut using a sliding window according to the user setting. BSMAP is used for the first-round mapping and remapping. We added '-S 123 -n 1 -r 0 -U' to the default BSMAP setting for uniquely mapped reads and reproducible results. Recombination of contiguously mapped fragments was implemented according to the following rules to reduce the possibility of false positives: 1) fragments must be strictly continuous, such that the distance between two fragments aligned on the genome must equal the distance between the two sequences on the read; 2) only one mismatch is allowed on each fragment; 3) if, after the recombination, a read overlaps two genomic fragments, the program will select the longest fragment with the fewest mismatches; and 4) all recombined fragments that do not exhibit read overlap are kept. To achieve better time and space complexity, the program records the number of mismatches on the last 'S' base pairs to accelerate the computation. 'S' stands for the length of the step. The sliding window in visualization module is achieved through "bedtools makewindows".

An HTML report template was provided in the pipeline. Datatable and JQuery were imported for data representation. Features of mapping and methylation calling were extracted from the original reports of the corresponding programs to a formatted text file for visualization. Figures were generated during the computation by the Matplotlib package in Python. Principle component analysis plot used methylation signals in sliding windows along the reference genome.

Simulation methods

The random nucleotide was generated by the random function in Numpy. The length of the real fragment in the middle was 110 bp from hg38. All cytosine were randomized as unmethylated or methylated. The chance of being methylated was equal to the methylation ratio of the corresponding CpG site in human embryonic stem cells. The random heads of reads had a random length (m) of 1–40 base pairs. The length of the random tail of each read was 40 (-m) base pairs. The genomic position of the middle part of the read was selected by a two-step randomization that generated the chromosome name and the starting point. For the fully randomized dataset, each nucleotide (A, C, T, or G) had an equal probability to fill each position.

WGBS data processing

All WGBS data mapped by first-round mapping with BSMAP are described as “without LiBis”, and WGBS data mapped by LiBis (i.e., clipped mapping followed by BSMAP), are described as “with LiBis”. scBS-map was applied with default parameters. LiBis used 40 as the window size, five as the stride, and 45 as the filter length. Picard(http://broadinstitute.github.io/picard) was used to remove PCR duplicates from both first round BSMAP results and second round LiBis rescued reads before analysis in processing clinical samples. The smoothed scatterplots (geneplotter in R package) used the CpGs sites common between the two samples as input. Pearson correlations were calculated using the R cor function. Boxplots were plotted using the Python package seaborn. Single cell whole genome bisulfite sequencing data is at the Gene Expression Omnibus database under accession number GSE56879. Data used to compute correlations between LiBis-specific CpGs and public tumor samples are MB tumor samples methylome at GSE54880 and GBM tumor samples methylome at GSE121721(19).

Results

LiBis: a highly cost-efficient, ultrasensitive alignment method for low-input bisulfite sequencing

LiBis is an integrated Python package for processing low-input WGBS data, such as cfDNA methylome and single-cell DNA methylome sequencing data. FastQC, Trim-galore, BSMAP, mcall, and bedtools are integrated into LiBis for quality control, adaptor trimming, read mapping, and methylation calling, respectively(15,20-22). The LiBis toolkit contains three modules: a preprocess module for quality control and adaptor trimming, a compute module for dynamic clipping and mapping, and a visualization module for report generation (**Figure 1A**, **Figure S1**).

To improve the mapping efficiency for low-input bisulfite sequencing data, we applied a clipping strategy within the compute module to eliminate random base contamination, as follows. First, LiBis maps the trimmed raw data with BSMAP and generates a new fastq file containing unmapped reads. Second, LiBis clips all unmapped reads using a sliding window of a specific width and step size as defined by the user. Third, LiBis remaps all clipped read fragments and keeps only uniquely mapped fragments for subsequent recombination. During recombination, fragments derived from the same unmapped read are recombined only if they are remapped contiguously to the reference genome, with the distance between two adjacent fragments being the step size. Recombined fragments are required to be above a minimum length to reduce the likelihood of a false-positive alignment. For reads with multiple candidate recombined clipped fragments that overlap, such as that illustrated in the top half of **Figure 1B**, the recombined clip with the greatest mapping confidence (i.e., the longest clip with least mismatches) is kept as a rescued read. If the recombined clipped fragments do not overlap each other, such as the two clips in the bottom half of **Figure 1B**, all recombined clipped fragments will be kept. Through the remapping and recombination steps mentioned above, reads discarded in first-round mapping can be rescued.

The LiBis workflow for bisulfite sequencing data involves five steps. In step 1, the raw reads are examined for quality control by FastQC. FastQC allows us to assess quality features including base quality, base content, and duplication level. These features reveal the overall quality of the library, amplification, and sequencing. Results of FastQC are aggregated in the final report. In step 2, the reads are trimmed by trim-galore, which removes sequencing adaptors and low-quality reads. If random priming is used to generate the library, trimming is recommended but will not remove random priming-associated amplification artifacts. Step 3 is read mapping by the compute module, which combines initial mapping with the dynamic clipping and remapping strategy to improve the cost efficiency. In step 4, the methylation ratio of CpGs is called by mcall within the MOABS program. In step 5, the data is visualized as various figures. After LiBis analysis, an overview webpage presents summaries of all input samples, heatmaps, and principal component analysis results (**Figure 1A**).

LiBis requires two types of input files, namely a fasta file containing the reference genome and the fastq files containing raw reads. For small numbers of samples, users can use command line to run the package. Additionally, we developed a config file reader for large number of samples. The LiBis output contains bam files, methylation ratio call files, quality control results, and an integrated HTML report.

Analysis of simulated data certified that LiBis provides enhanced mapping ratios, sensitivity, and efficiency

To validate the ability of LiBis on rescuing missed reads by traditional bisulfite alignment tools, simulation dataset including 3 samples with 20 million 150 base pair reads was generated *in silico*. For each sample, 10 million reads were fully random in which every single base had equal changes of being A, T, C or G by which the false positive rate of LiBis is proved. Each of the other half of the reads was composed of 110-bp DNA fragment (chosen randomly from the hg38 reference genome) flanked by random sequences on both ends as random head and random tail. The random heads and tails represented the contamination introduced by random priming, which is not fully removed by traditional trimming

methods such as trim-galore. These reads with “contamination” were used to test the positive discovery rate(**Figure 2A**).

The minimal length of rescued reads by LiBis relates to mapping accuracy. We investigated the true positive rates with different filter length settings. As shown in Figure 2B, by selecting 45 base pair as the filter length, the true positive rate was saturated.. All simulated samples were mapped using LiBis and scBS-map as the state of the art(18). Regarding detectability, LiBis mapped an average of 92.6% of the reads among the three read samples and achieved a higher discovery rate than scBS-map (**Figure 2C**). Regarding sensitivity, an event was deemed to be a true positive if the interval of the simulated read on the genome included its mapping interval. For the mapped reads, LiBis achieved over 99.9% true positive rate on average, which was significantly higher than the 92.9% true positive rate of scBS-map (**Figure S3**). The CpGs identified by LiBis also showed a high correlation (R=0.98) with the ground truth for the simulation (**Figure 2D**), indicating the accuracy of LiBis. These results revealed that LiBis achieves high detectability and high sensitivity in simulations.

To further investigate the efficiency of LiBis, we recorded the average CPU time of LiBis for processing simulated samples. Our result showed that LiBis used only 20.9 CPU hours, which was nearly four times faster than scBS-map (**Figure 2E**).

LiBis improved the cost efficiency and accuracy of single-cell methylome profiling

To test the efficiency of LiBis for real-world data, we applied LiBis to a public scWGBS dataset (GSE56879). We observed that the mapping ratio for the scWGBS samples increased by as much as 24% when LiBis was applied as opposed to BSMAP (**Figure 3A**). To measure the effect of this improvement on downstream analysis at a single-CpG level, we defined informative CpGs as those with coverage ≥5 reads. For this scWGBS dataset, LiBis increased the number of informative CpGs by as much as 70% compared to BSMAP, which significantly enlarged the pool of usable CpG sites for identifying

differentially methylated CpGs. Moreover, LiBis also identified new CpG sites which had no coverage in the first-round mapping in any sample (**Figure 3C**). The CpG sites specifically discovered with LiBis had relatively high methylation ratios compared to the initially discovered CpGs (**Figure S3A**), indicating that reads from highly methylated regions may suffer more contamination from random priming. Because clipped reads are not usually discovered by typical alignment programs, our result may explain the lower methylation ratios detected in libraries built using random priming versus in libraries prepared using other methods(23).

To further investigate the influence of LiBis on methylome measurement precision, we compared the scWGBS methylome processed by LiBis to the corresponding bulk methylome processed by BSMAP. We assumed the methylation ratios of the deeply measured CpG sites in the bulk sample processed by BSMAP to be the gold standard if we followed typical processing procedures, though the gold standard may not be ground truth. For each bulk tissue, we merged all matched single-cell samples processed by LiBis with equal weight and obtained a merged methylome. There were 2.5 million CpGs in common between the LiBis merged methylome and the matched BSMAP bulk methylome, and we observed a high correlation between the two methods at these CpGs (**Figure 3D**). In gene regions, the LiBis merged methylome and the matched BSMAP bulk methylome showed nearly identical methylation ratio distributions (**Figure S3B**). This revealed that applying LiBis does not change the methylation ratio distribution for most CpG sites. Of the 2.5 million common CpGs, 187,804 were specifically identified by LiBis in the scWGBS data; the correlation for these LiBis-specific CpGs between the matched BSMAP bulk methylome and the LiBis merged methylome(R=0.952) (**Figure 3E**) was higher than that for all common CpGs (**Figure 3D**). Additionally, the absolute methylation ratio difference between the BSMAP bulk methylome and the merged scWGBS methylome by BSMAP or LiBis was significantly lower (p-value <0.01) when BSMAP was replaced with LiBis (**Figure 3F**). These results revealed that the rescued reads and CpGs are beneficial for downstream analysis and that the methylation information retrieved by LiBis is highly accurate.

LiBis improved the data efficiency of tumor and cfDNA WGBS experiments with random priming

To verify the capability of LiBis for rescuing bisulfite sequencing from cfDNA, we performed WGBS on six clinical samples and processed the sequencing data by LiBis: 1). Two cerebrospinal fluid(CSF) cfDNA samples from pediatric medulloblastoma patients; 2). Two cfDNA samples from glioblastoma (GBM) patients' plasma; 3). Two genomic DNA samples from GBM tumor tissues (**Table S1**). Previous studies reported that CSF has very low cfDNA content that leads to low mapping ratios and insufficient coverage in sequencing analyses(24). Like is done in scWGBS to utilize as much of the DNA content as possible, bases were randomly assigned to the heads and tails of bisulfite-treated fragments. By analyzing the base content of the sequences, we found that our samples prepared using the random priming library method showed a strong C-bias at the beginning of the reads. This observation was also reported during single-cell bisulfite sequencing method development due to the dNTP in random priming [14]. Trim-galore could not eliminate the C-bias, indicating that random artifactual bases cannot be removed by traditional trimming methods (**Figure S4A**). After removing the artifactual bases, we observed that the fragment length of LiBis rescued reads is slightly lower than common cfDNA fragment length(Figure S4B).

Through dynamic clipping, LiBis significantly increased the mapping ratio by average 71.92%(range from 50.07%-88.53%) among clinical samples (**Figure 4A**). The number of CpG sites that were only covered by LiBis rescued reads for each dataset is shown in **Figure 4B**.We observed high correlation coefficient (average 0.85, 0.75-0.92) between the first round BSMAP mapping and the second round clipped mapping at common CpGs covered least 10 times (**Figure 4C**, **Figure S4C).** We further investigated the methylation difference between these common CpGs to validate the consistency between the first round BSMAP mapping and the clipped mapping. For every sample, over 90% CpG sites have no significant differences (methylation ratio difference<0.2) (**Figure 4D**). These results proved that DNA methylation ratios of the common CpGs detected by both BSMAP and LiBis are highly consistent. For those CpGs that specifically rescued by LiBis reads, the correlation coefficient with public tumor tissue methylome reached 0.8 for genomic DNA and 0.7 for cell free DNA (**Figure 4E**). These results strongly

proved that LiBis can efficiently extract methylation information from low-input bisulfite sequencing data of clinical samples.

LiBis identified human papillomavirus integration sites in a tumor using WGBS with a traditional protocol

To further explore the capability of LiBis to discover genomic patterns, we collected one cervical cancer tumor sample and performed WGBS using a traditional protocol to prepare the sequencing library. The mapping ratio improved from 84% with BSMAP to 90% with LiBis for the 150 bp paired-end sequencing. Although the improvement was relatively small compared to that observed for the cfDNA WGBS data, we confirmed that WGBS together with LiBis could identify DNA fusion sites, in particular the human papillomavirus(HPV) DNA-human DNA fusion sites in this case study. In cervical cancer, HPV DNA can cleave human DNA and insert itself into the human DNA in a sense or antisense direction. Such fusion reads containing both HPV and human DNA fragments were identified by LiBis, which informs on the insertion spots of HPV (**Figure 5A**). We identified a list of potential insertion sites in this patient sample using LiBis (**Figure 5B**). For each insertion site, we identified the closest gene to find the potential influence of the virus insertion on gene expression. Interestingly, four of six genes were previously reported as HPV insertion sites or were differentially expressed in cervical cancer(25-27). As examples, one insertion site was located in the intron of the *CASC11* gene and in the flanking promoter region of the *MYC* gene (**Figure 5C**). *CASC11*, a gene that is upregulated in cervical cancer tissue, might activate the WNT/β-catenin signaling pathway to promote cervical cancer, and the *MYC* gene is well known for contributing to tumorigenesis(28). Interestingly, at this insertion site HPV integrates in a non-linear fashion. In addition to the identification of the HPV integration sites, LiBis enabled the identification of differentially methylated CpGs in both human and HPV genomes using the same pipeline.

Discussion

Liquid biopsy using cfDNA, as opposed to tissue biopsy, has outstanding advantages such as comprehensive representation of tumors, and noninvasive, repeatable, and quick sampling. However, the low concentration of cfDNA in plasma or CSF makes it difficult to analyze. This issue often results in low mapping ratios and, consequently, insufficient site coverage and low sequencing depth across the genome, especially in WGBS. Bisulfite conversion degrades the small cfDNA fragments, which leads to increased noise in deep sequencing. The cfDNA fragments peak near only 120 base pairs before bisulfite damage, but the most economic sequencing method is the 150 bp paired-end sequencing mode on the Illumina NovaSeq platform. Developing a computational method to improve the mapping efficiency and thus the data accuracy of low-input bisulfite sequencing has become a prerequisite to identify methylation biomarkers from cfDNA.

LiBis was capable of improving the cost efficiency of the experiments and the measurement precision for all tested datasets. We confirmed that LiBis improved performance in four case studies, namely simulated data, scWGBS data, cfDNA WGBS data from a random priming library, and tumor WGBS data from a traditional library. The performance improvement may due to the following reasons. First, random priming, adapter dimers, small fragments, or bisulfite conversion may generate artifactual bases at both ends of the reads. Second, through dynamic clipping, LiBis could recognize DNA fusions present in the original cell, including gene fusions, insertions of HPV DNA, fusions that arose during the library preparation, and even circular DNA formed by self-priming or other mechanisms.

To the best of our knowledge, LiBis is the first integrated pipeline for the processing of low-input bisulfite sequencing data that includes trimming, initial mapping, clipped mapping, methylation calling, and visualization. LiBis can be deployed easily on local computers, computer clusters, or cloud networks. Although our strategy improved the sequencing output: cost ratio for low-input bisulfite sequencing, there is additional computation required compared to the typical method.

Conclusions

We developed LiBis, a novel dynamic clipping and mapping method for WGBS of low-input DNA. Using fastq files as input, LiBis outputs a well-organized HTML report of results from the different modules. LiBis significantly improves the cost efficiency for analyzing low-input bisulfite sequencing data by providing a larger number of accurate informative CpGs and increasing the sequencing depth of all CpGs for downstream analysis. LiBis provides accurate methylome information through conservative mapping strategies, as illustrated by the simulation data and the data from scWGBS and bulk WGBS experiments. Moreover, LiBis can identify DNA fusion features, such as virus insertion sites, from WGBS data.

Data Access

All raw and processed sequencing data generated in this study are available at the Gene Expression Omnibus database under accession number GSE142241.

Software Requirements

Project name: LiBis

Operating system(s): Linux or system based on Docker environment

Programming language: Python 3.6

Other requirements: Docker (Optional)

License: GNU

List of Abbreviations

LiBis: low-input bisulfite sequencing alignment

cfDNA: cell-free DNA

WGBS: whole-genome bisulfite sequencing

CpG: cytosine nucleotide followed by a guanine nucleotide

ctDNA: circulating tumor DNA

HTML: HyperText Markup Language

HPV: Human papillomavirus

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by award RP180131 from the Cancer Prevention and Research Institute of Texas and by start-up funds from Texas A&M University.

Acknowledgements

We are grateful to the Texas A&M University Brazos HPC cluster (brazos.tamu.edu) and the Texas A&M Institute for Genome Sciences and Society (TIGSS) HPC cluster (tigss.tamu.edu) that contributed to the research reported here.

Author Contributions

DS directed and oversaw the project. YY developed and tested the pipeline and performed the computational experiments. YY and JL performed comprehensive bioinformatics analysis, including data quality control, publicly available data collection, and integration analysis. ML and YH optimized CSF ctDNA sequencing library preparation and performed high-throughput sequencing. SZ, XNL, and LG collected the samples. Jin, Jianfang, and Mutian provided intellectual input. YY and DS wrote the manuscript. All authors contributed to discussion, program testing, and the writing of the manuscript.

References

1. Mouliere, F. and Rosenfeld, N. (2015) Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc Natl Acad Sci U S A*, **112**, 3178-3179.
2. Satouchi, M., Tanaka, H., Yoshioka, H., Shimokawaji, T., Mizuno, K., Takeda, K., Yoshino, I., Seto, T., Kurata, T. and Tashiro, N. (2017) Detection of epidermal growth factor receptor gene T790M mutation in cytology samples using the cobas® EGFR mutation test. *Lung Cancer*, **111**, 190-194.
3. Lamb, Y.N. and Dhillon, S. (2017) Epi proColon® 2.0 CE: a blood-based screening test for colorectal cancer. *Molecular diagnosis & therapy*, **21**, 225-232.
4. Underhill, H.R., Kitzman, J.O., Hellwig, S., Welker, N.C., Daza, R., Baker, D.N., Gligorich, K.M., Rostomily, R.C., Bronner, M.P. and Shendure, J. (2016) Fragment length of circulating tumor DNA. *PLoS genetics*, **12**.
5. Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C., Modlin, L.A., Liu, C.L., Neal, J.W., Wakelee, H.A., Merritt, R.E. *et al.* (2014) An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med*, **20**, 548-554.
6. Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*, **28**, 1097-1105.
7. Warnecke, P.M., Stirzaker, C., Song, J., Grunau, C., Melki, J.R. and Clark, S.J. (2002) Identification and resolution of artifacts in bisulfite sequencing. *Methods*, **27**, 101-107.
8. Urich, M.A., Nery, J.R., Lister, R., Schmitz, R.J. and Ecker, J.R. (2015) MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc*, **10**, 475-483.
9. Warnecke, P.M., Stirzaker, C., Melki, J.R., Millar, D.S., Paul, C.L. and Clark, S.J. (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res*, **25**, 4422-4426.
10. Genereux, D.P., Johnson, W.C., Burden, A.F., Stoger, R. and Laird, C.D. (2008) Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res*, **36**, e150.
11. Miura, F., Enomoto, Y., Dairiki, R. and Ito, T. (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res*, **40**, e136.
12. Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J. *et al.* (2018) Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol*, **36**, 428-431.
13. Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W. and Kelsey, G. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods*, **11**, 817-820.
14. Luo, C., Rivkin, A., Zhou, J., Sandoval, J.P., Kurihara, L., Lucero, J., Castanon, R., Nery, J.R., Pinto-Duarte, A., Bui, B. *et al.* (2018) Robust single-cell DNA methylome profiling with snmC-seq2. *Nat Commun*, **9**, 3824.

15. Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A. and Li, W. (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol*, **15**, R38.
16. Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571-1572.
17. Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.-Y. and Pellegrini, M. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC genomics*, **14**, 774.
18. Wu, P., Gao, Y., Guo, W. and Zhu, P. (2019) Using local alignment to enhance single-cell bisulfite sequencing data efficiency. *Bioinformatics*, **35**, 3273-3278.
19. Hovestadt, V., Jones, D.T., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski, K., Ryzhova, M. and Warnatz, H.-J. (2014) Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*, **510**, 537-541.
20. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
21. Xi, Y. and Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.
22. Bioinformatics, B. (2011) FastQC: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute*.
23. Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R.V., Branco, M.R. and Reik, W. (2018) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol*, **19**, 33.
24. Pan, W., Gu, W., Nagpal, S., Gephart, M.H. and Quake, S.R. (2015) Brain tumor mutations detected in cerebral spinal fluid. *Clinical chemistry*, **61**, 514-522.
25. Holmes, A., Lameiras, S., Jeannot, E., Marie, Y., Castera, L., Sastre-Garau, X. and Nicolas, A. (2016) Mechanistic signatures of HPV insertions in cervical carcinomas. *NPJ genomic medicine*, **1**, 16004.
26. Li, W., Tian, S., Wang, P., Zang, Y., Chen, X., Yao, Y. and Li, W. (2019) The characteristics of HPV integration in cervical intraepithelial cells. *Journal of Cancer*, **10**, 2783.
27. Tripathi, R., Rath, G., Hussain, S., Jawanjal, P., Bandil, K., Sharma, V., Bharadwaj, M. and Mehrotra, R. (2018) Jagged-1 induced molecular alterations in HPV associated invasive squamous cell and adenocarcinoma of the human uterine cervix. *Scientific reports*, **8**, 1-9.
28. Hsu, W., Liu, L., Chen, X., Zhang, Y. and Zhu, W. (2019) LncRNA CASC11 promotes the cervical cancer progression by activating Wnt/beta-catenin signaling pathway. *Biological research*, **52**, 33.

Figure Legends

Figure 1. Development of LiBis (low-input bisulfite sequencing alignment). A. LiBis overview. B. Details of the LiBis rescue procedure, including clipping initially unmapped reads, remapping clipped fragments, and recombining contiguous fragments.

Figure 2. LiBis shows a high discovery rate, high sensitivity, high accuracy, and fast efficiency *in silico*. A. Scheme for generating simulated reads with random heads and tails. B. True positive rates of LiBis in different filter lengths. C. Mapping ratios of simulated reads by LiBis or scBS-map. D. Methylation ratio correlation between simulation ground truth and LiBis. E. CPU times for processing 10 million reads by LiBis or scBS-map.

Figure 3. LiBis improves both the efficiency and the precision of methylation measurements in scWGBS samples. A–C. Percentage change for the mapping ratio, number of informative CpGs (depth ≥5), and CpGs covered only by LiBis-rescued reads between LiBis and BSMAP. D. Correlation of the methylation ratio (averaged for 1-kb windows) between bulk and merged methylomes. E. Methylation ratio correlation between bulk and merged methylomes at regions specifically discovered by LiBis. F. The difference in absolute methylation ratio between bulk and merged methylomes is reduced after LiBis.

Figure 4. LiBis improves the cost efficiency of clinical tumor and cfDNA WGBS data using the random priming method. A. Mapping ratio improvement by LiBis for clinical samples. B. Number of CpGs specifically discovered by LiBis in clinical samples. C. Correlation of the methylation ratios between the first round BSMAP mapping and the second round clipped mapping. The correlation coefficients are listed on the top of the figures. D. Insignificant methylation differences on common CpGs between two round mapping processes. E. Correlations of methylation ratios between informatics LiBis-specific CpGs and collected public patients' methylome.

Figure 5. LiBis identifies HPV insertion sites in cervical cancer WGBS experiments using a traditional pre-bisulfite library method. A. Scheme of HPV insertion. B. HPV insertion sites in cervical cancer identified by LiBis. C. Visualization of an identified HPV insertion site.

Supplementary Figure Legends

Figure S1. Example of the final report generated by the LiBis visualization module. Top panel, table of basic statistics and HTML links to FastQC figures. Bottom panel, principle component analysis scatter plot and heatmap.

Figure S2. True positive rates of mapped reads by LiBis or scBS-map.

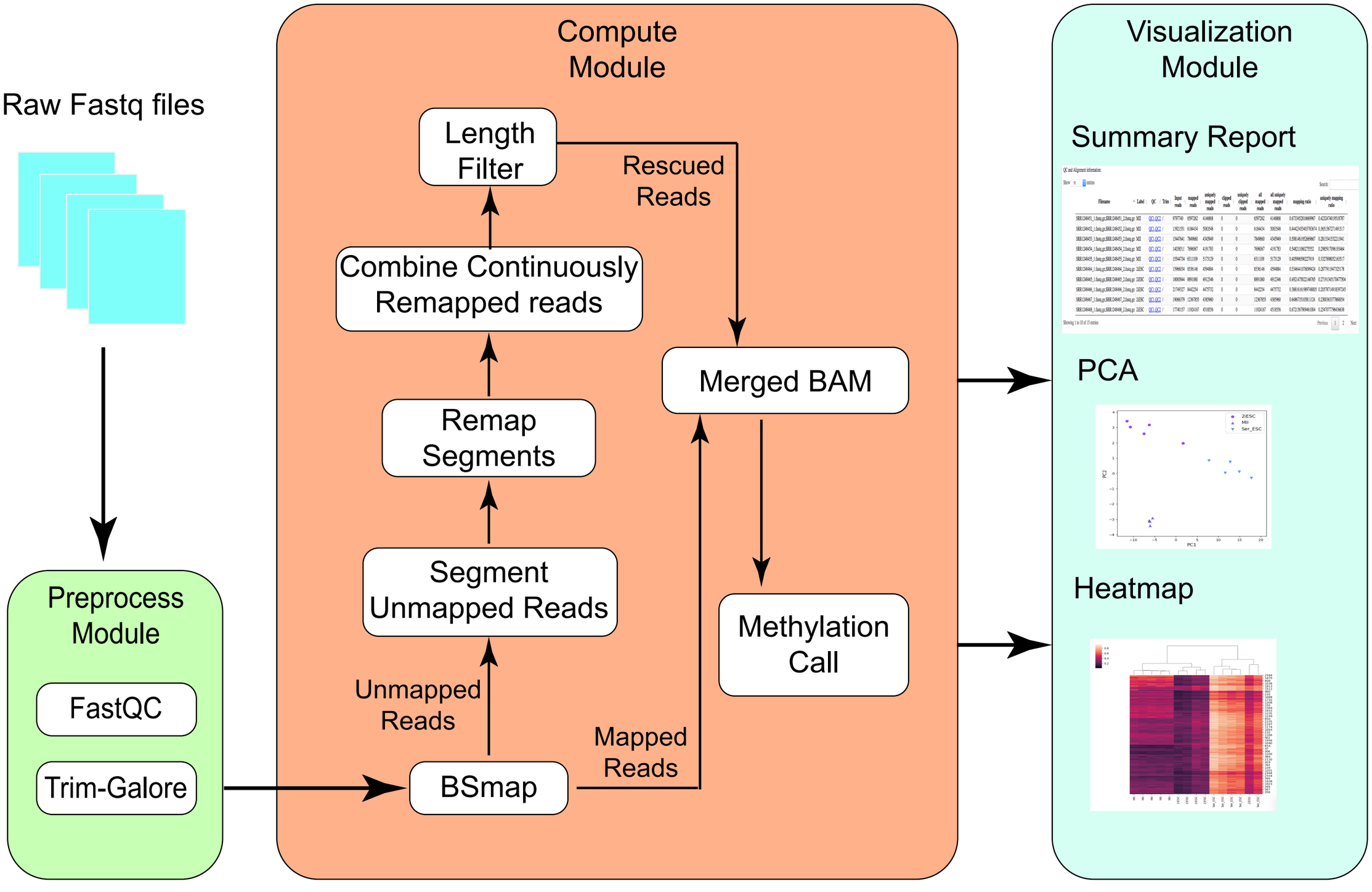
Figure S3. A. Regions specifically discovered by LiBis have a higher methylation ratio than other regions. B. Merged methylomes generated with LiBis or without LiBis exhibit nearly identical methylation ratio distributions over gene regions. All genes in the genome were stretched to 1 kb and aligned to the interval from Start to End on the X axis. The Y axis is the methylation ratio. The bulk methylome by BSMAP is in red, the merged methylome with or without LiBis is in green or blue, respectively.

Figure S4. A. Bases content distribution with or without LiBis. Content bias cannot be eliminated by traditional trimming methods, although it can be removed by LiBis. LiBis has stronger ability of eliminating bases content bias than traditional methods. B. Fragment length distribution of LiBis rescued reads. C. Correlation of the methylation ratios between the first round BSMAP mapping and the second round LiBis clipped mapping.

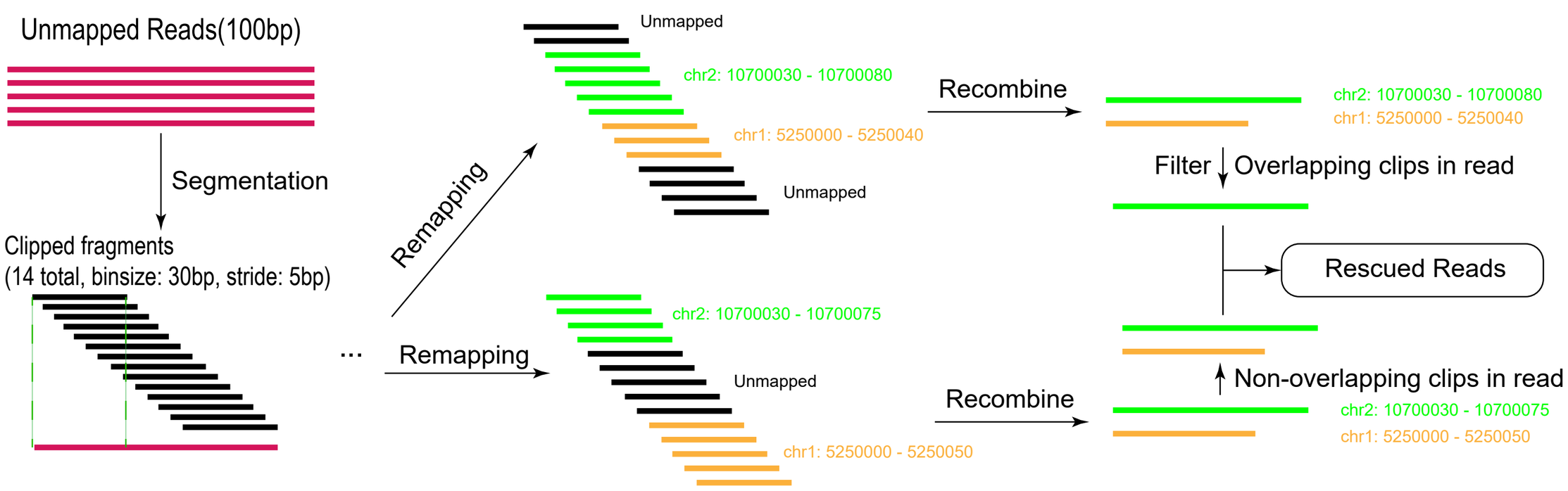
Supplementary Table Legend

Table S1. Sequencing data statistics

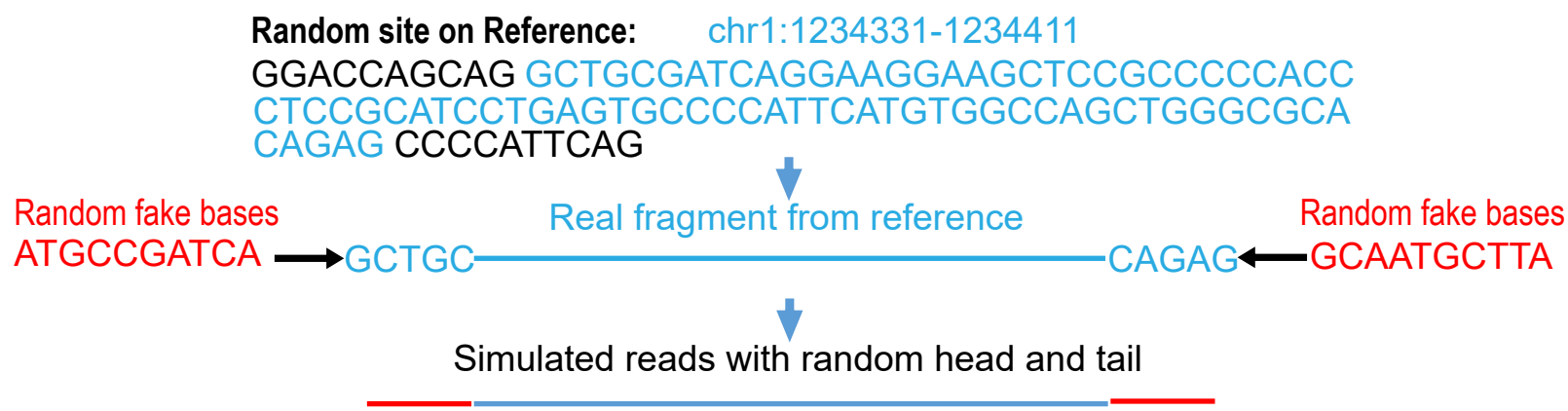
A


Raw Fastq files
Preprocess Module
FastQC
Trim-Galore
Compute Module
Length Filter
Rescued Reads
Combine Continuously Remapped reads
Remap Segments
Segment Unmapped Reads
Unmapped Reads
BSmap
Mapped Reads
Merged BAM
Methylation Call
Visualization Module
Summary Report
PCA
Heatmap

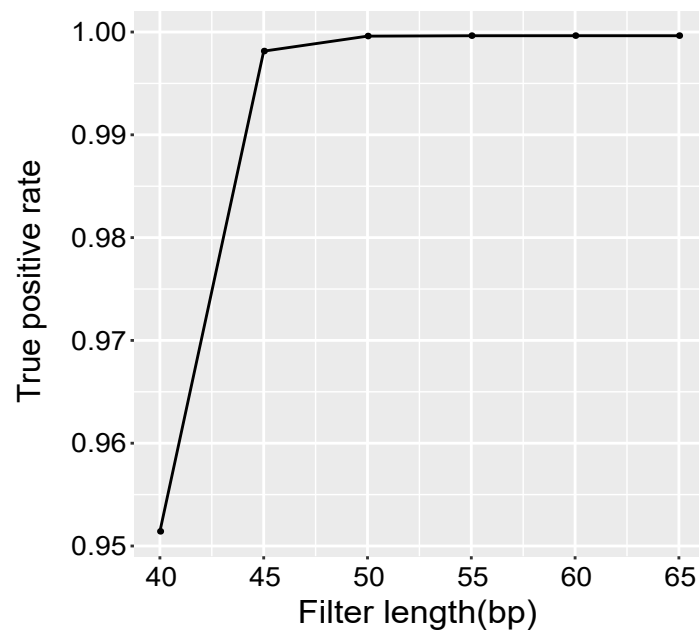
B


Unmapped Reads(100bp)
Segmentation
Clipped fragments
(14 total, binsize: 30bp, stride: 5bp)
Remapping
Remapping
Unmapped
chr2: 10700030 - 10700080
chr1: 5250000 - 5250040
Unmapped
chr2: 10700030 - 10700075
Unmapped
chr1: 5250000 - 5250050
Recombine
Recombine
chr2: 10700030 - 10700080
chr1: 5250000 - 5250040
Filter
Overlapping clips in read
Rescued Reads
Non-overlapping clips in read
chr2: 10700030 - 10700075
chr1: 5250000 - 5250050

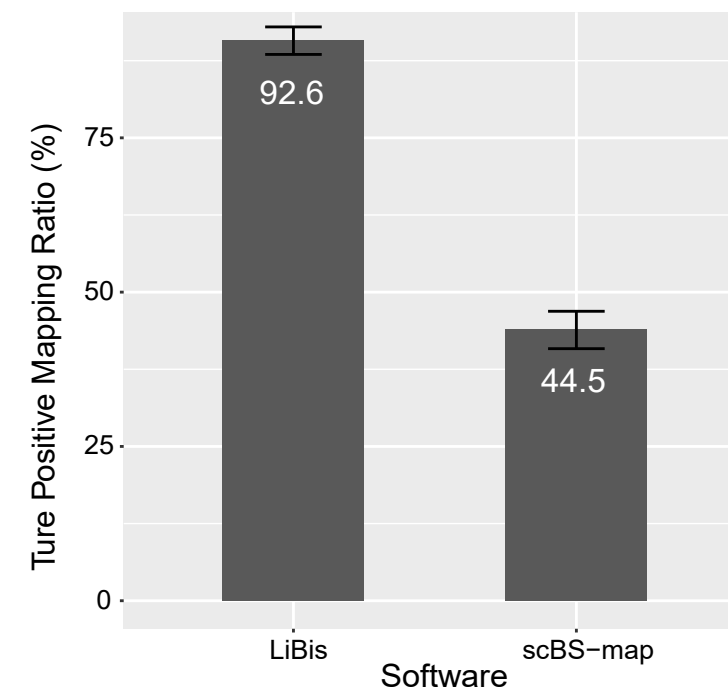
A


Random site on Reference:
chr1:1234331-1234411
GGACCAGCAG GCTGCGATCAGGAAGGAAGCTCCGCCCCCACC
CTCCGCATCCTGAGTGCCCCATTCATGTGGCCAGCTGGGCGCA
CAGAG CCCCATTCAG
Random fake bases
ATGCCGATCA
GCTGC
Real fragment from reference
CAGAG
Random fake bases
GCAATGCTTA
Simulated reads with random head and tail

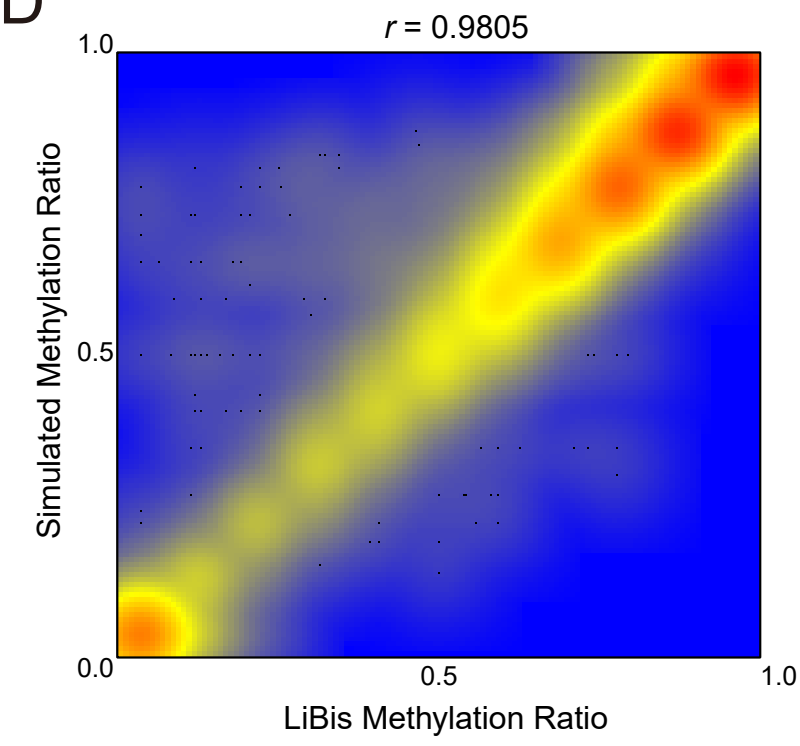
B


True positive rate
1.00
0.99
0.98
0.97
0.96
0.95
40
45
50
55
60
65
Filter length(bp)

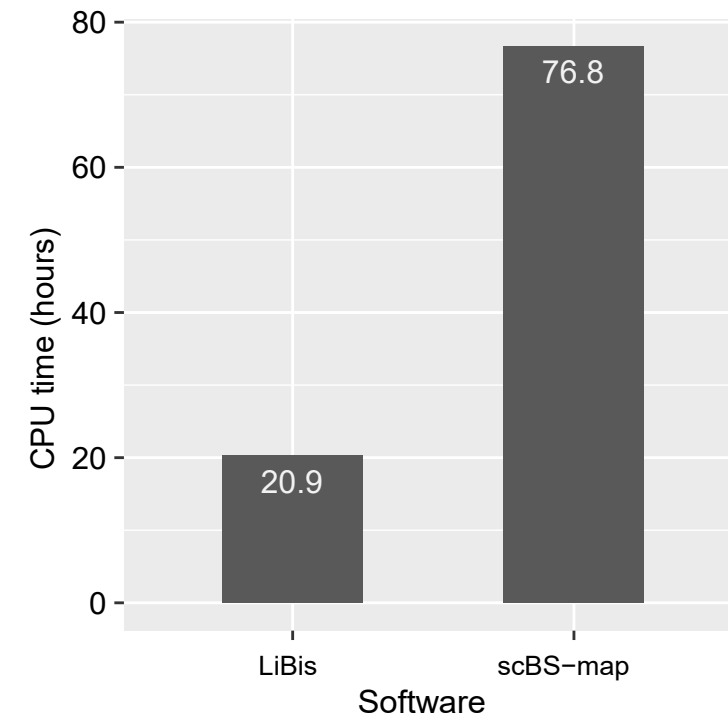
C

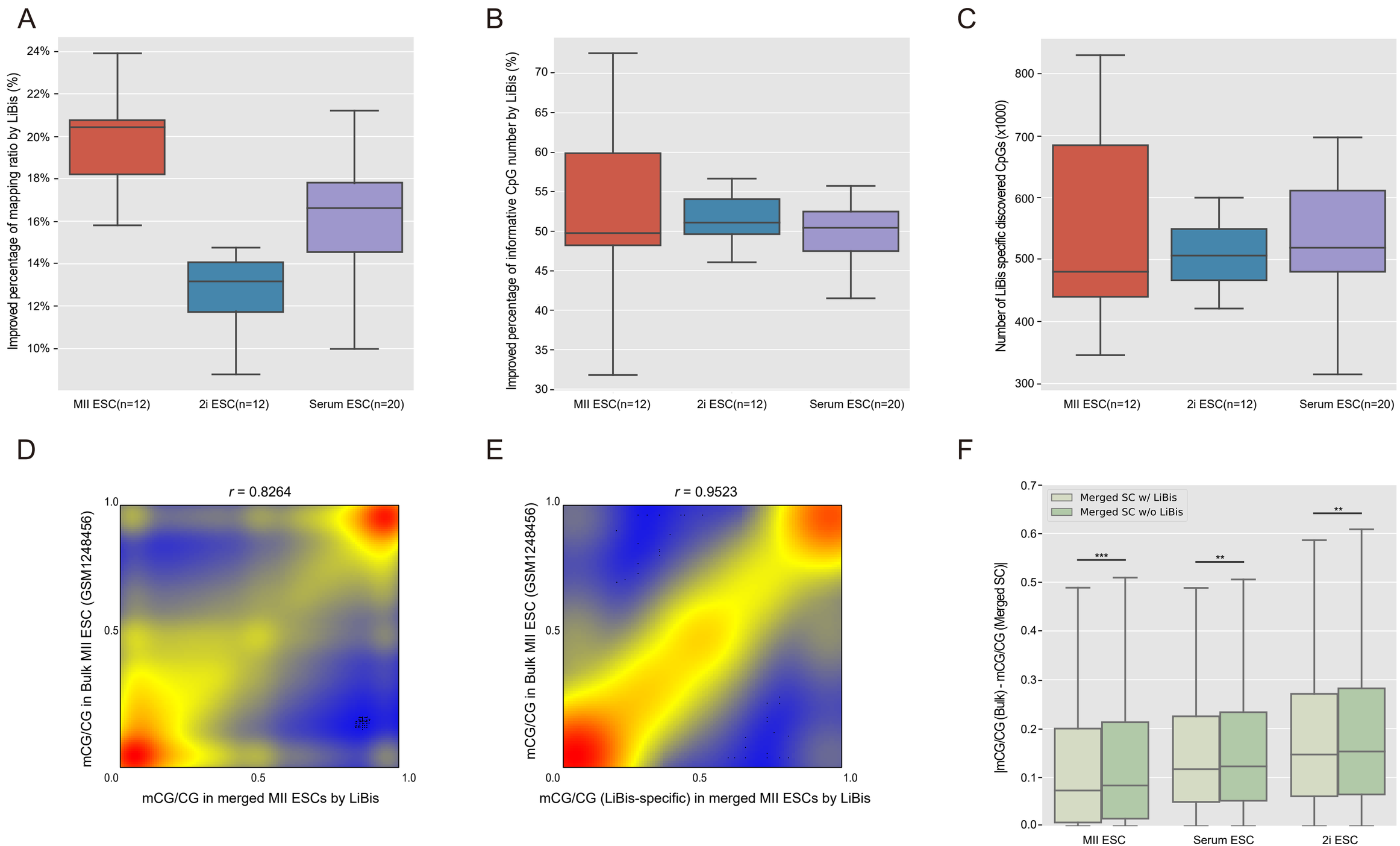

Ture Positive Mapping Ratio (%)
75
50
25
0
92.6
44.5
LiBis
scBS−map
Software

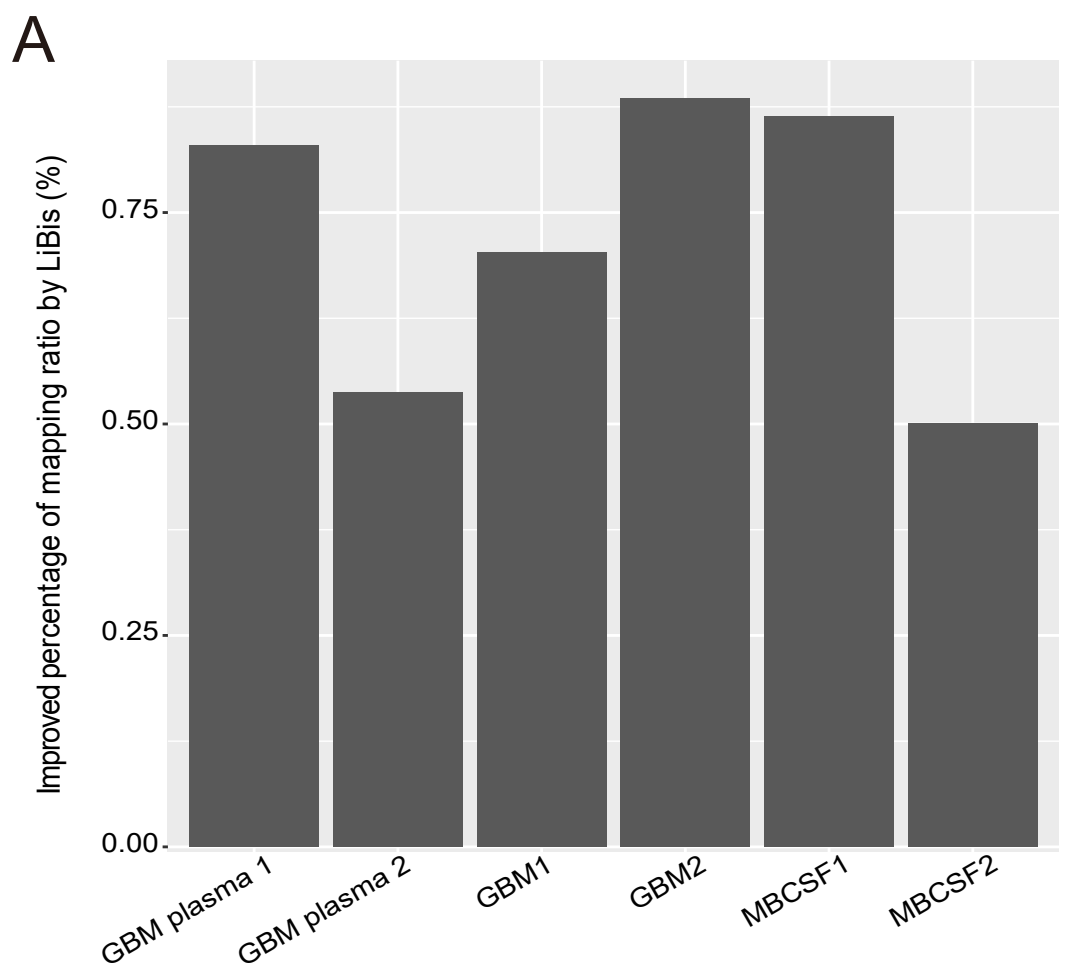
D

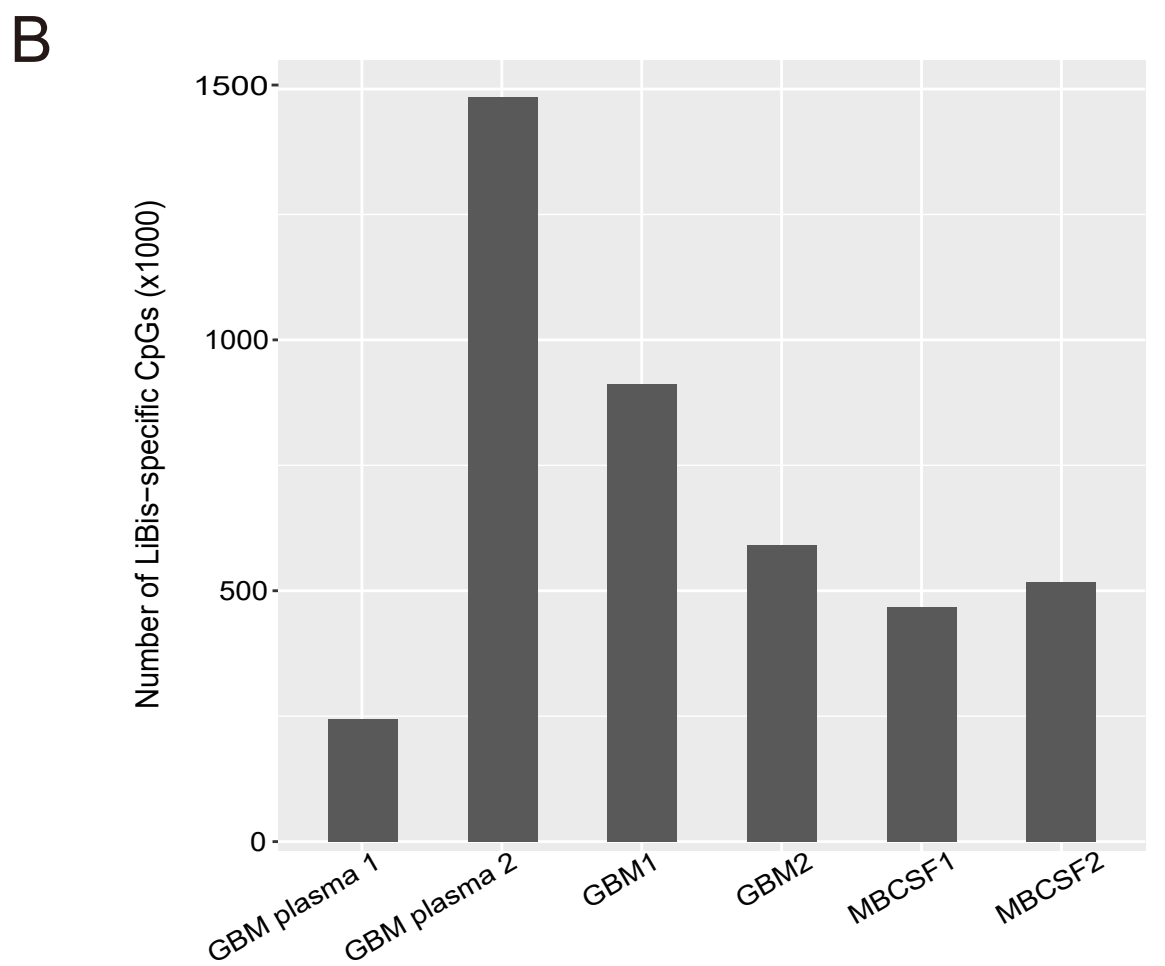

r = 0.9805
Simulated Methylation Ratio
1.0
0.5
0.0
0.5
1.0
LiBis Methylation Ratio

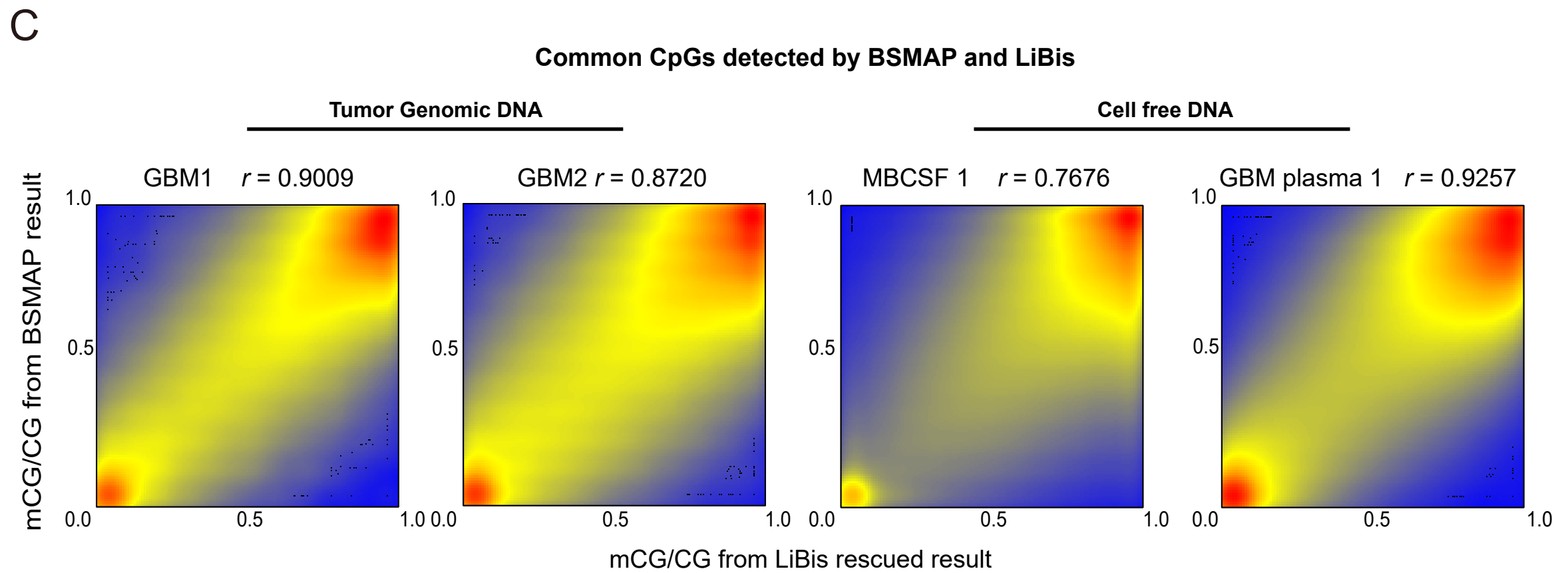
E

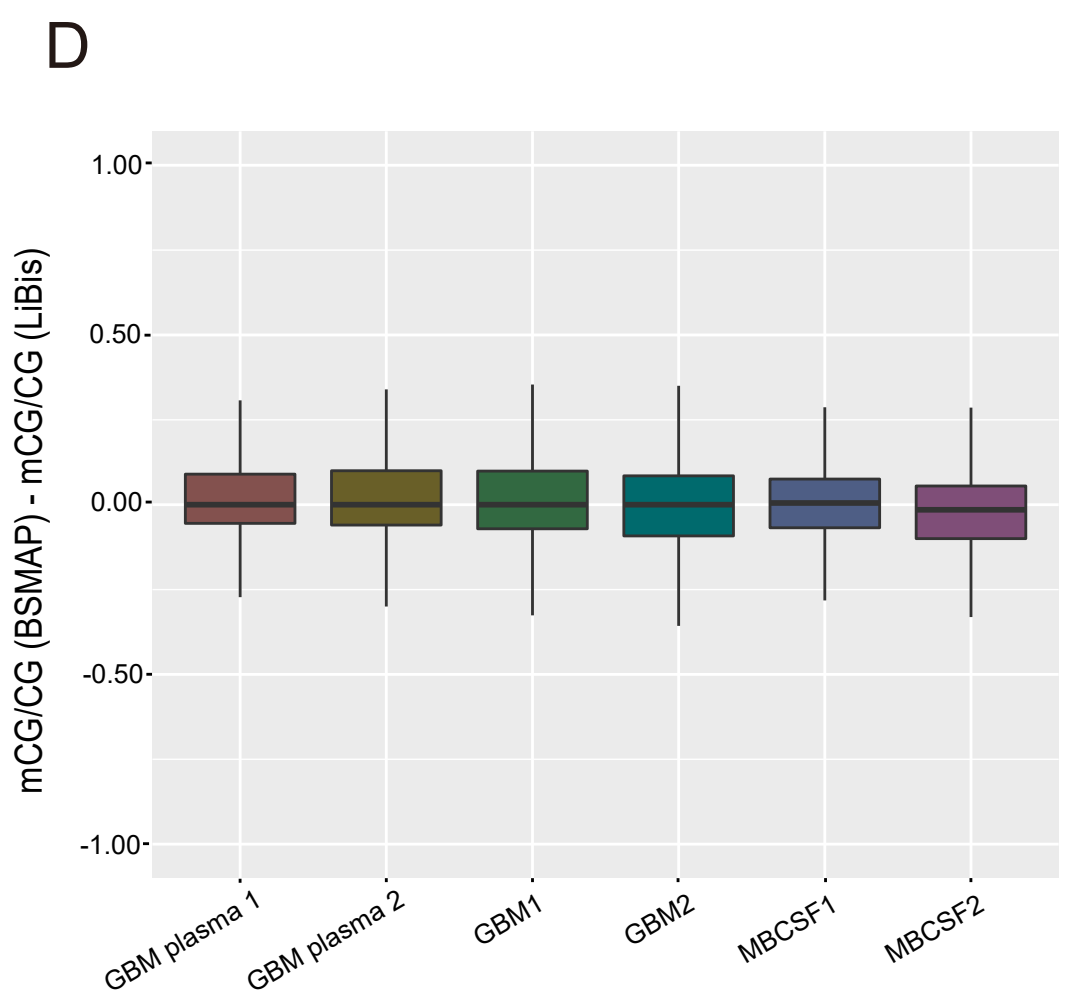

CPU time (hours)
80
60
40
20
0
76.8
20.9
LiBis
scBS−map
Software

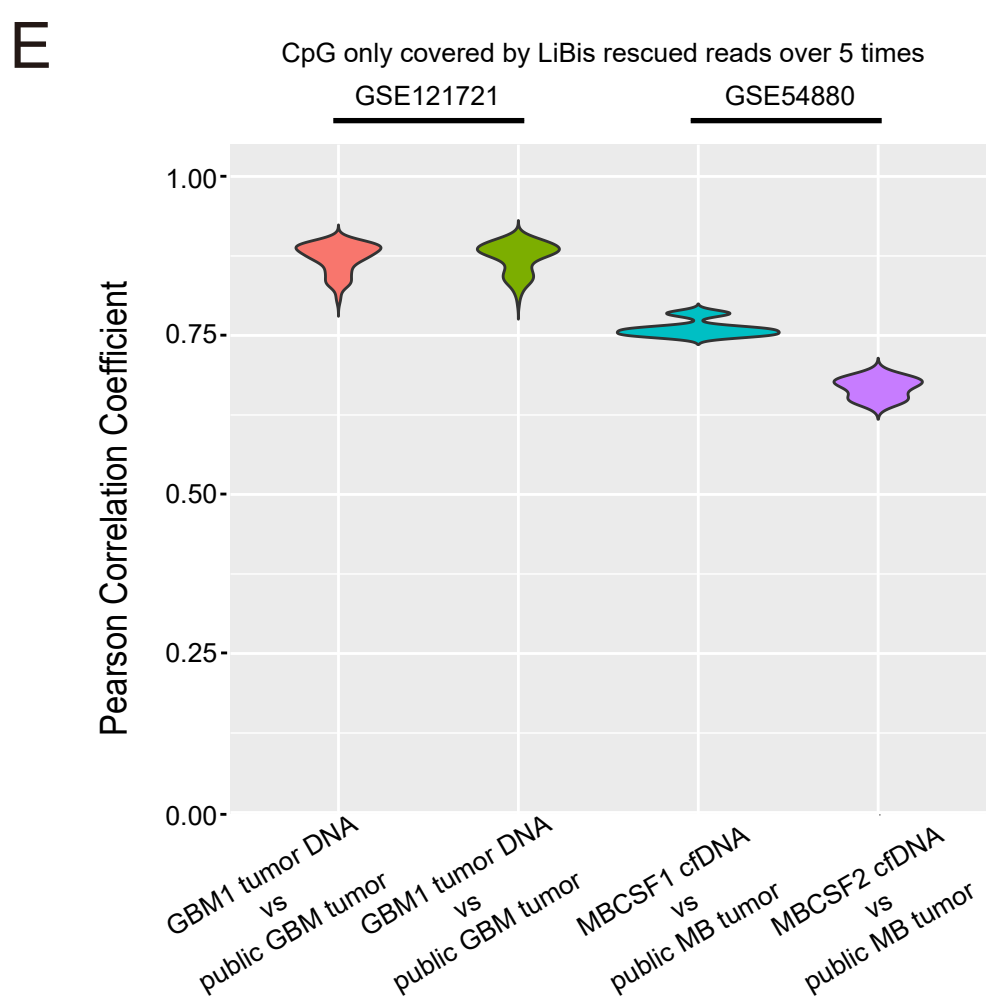

A
Improved percentage of mapping ratio by LiBis (%)
24%
22%
20%
18%
16%
14%
12%
10%
MII ESC(n=12)
2i ESC(n=12)
Serum ESC(n=20)
B
Improved percentage of informative CpG number by LiBis (%)
70
65
60
55
50
45
40
35
30
MII ESC(n=12)
2i ESC(n=12)
Serum ESC(n=20)
C
Number of LiBis specific discovered CpGs (x1000)
800
700
600
500
400
300
MII ESC(n=12)
2i ESC(n=12)
Serum ESC(n=20)
D
r = 0.8264
mCG/CG in Bulk MII ESC (GSM1248456)
1.0
0.5
0.0
0.5
1.0
mCG/CG in merged MII ESCs by LiBis
E
r = 0.9523
mCG/CG in Bulk MII ESC (GSM1248456)
1.0
0.5
0.0
0.5
1.0
mCG/CG (LiBis-specific) in merged MII ESCs by LiBis
F
Merged SC w/ LiBis
Merged SC w/o LiBis
|mCG/CG (Bulk) - mCG/CG (Merged SC)|
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
***
**
**
MII ESC
Serum ESC
2i ESC


A
Improved percentage of mapping ratio by LiBis (%)
0.00
0.25
0.50
0.75
GBM plasma 1
GBM plasma 2
GBM1
GBM2
MBCSF1
MBCSF2

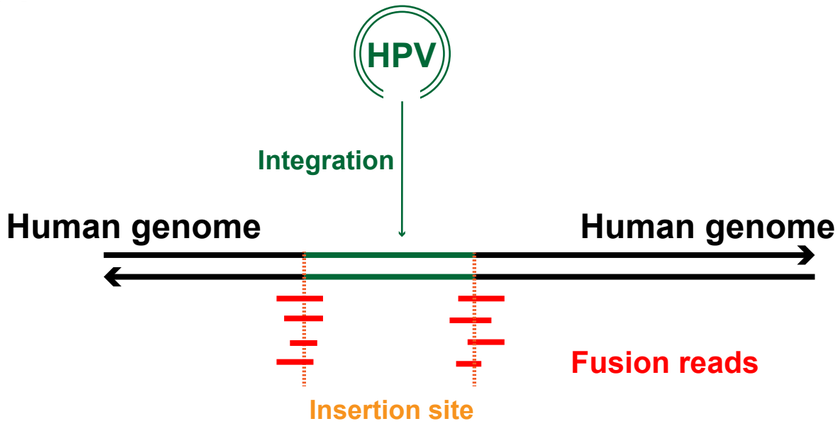

B
Number of LiBis−specific CpGs (x1000)
0
500
1000
1500
GBM plasma 1
GBM plasma 2
GBM1
GBM2
MBCSF1
MBCSF2


C
Common CpGs detected by BSMAP and LiBis
Tumor Genomic DNA
Cell free DNA
GBM1 r = 0.9009
GBM2 r = 0.8720
MBCSF 1 r = 0.7676
GBM plasma 1 r = 0.9257
mCG/CG from BSMAP result
mCG/CG from LiBis rescued result
0.0
0.5
1.0

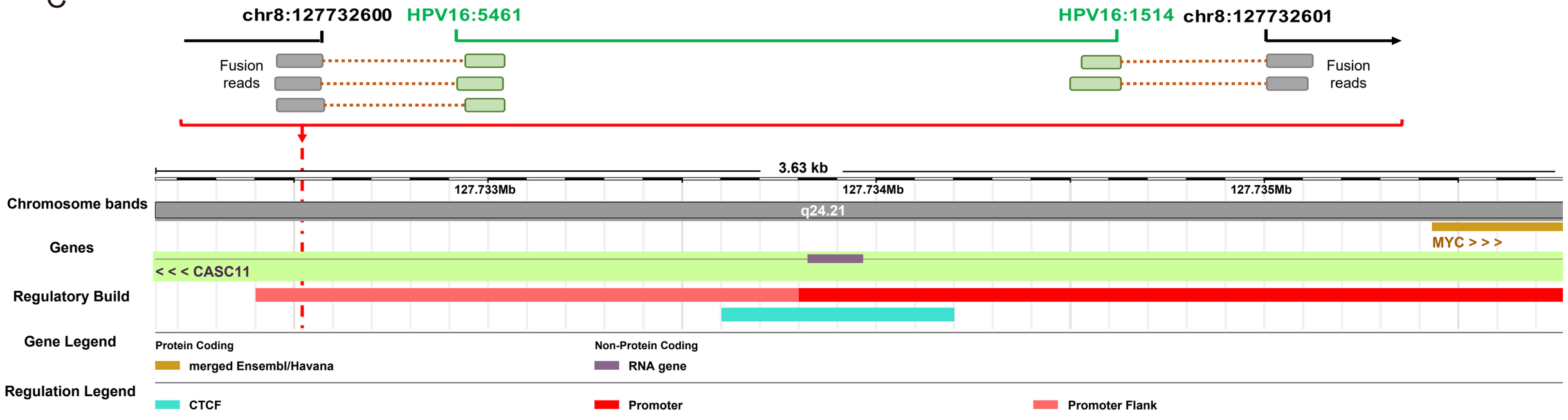

D
mCG/CG (BSMAP) - mCG/CG (LiBis)
1.00
0.50
0.00
-0.50
-1.00
GBM plasma 1
GBM plasma 2
GBM1
GBM2
MBCSF1
MBCSF2


E
CpG only covered by LiBis rescued reads over 5 times
GSE121721
GSE54880
Pearson Correlation Coefficient
0.00
0.25
0.50
0.75
1.00
GBM1 tumor DNA vs public GBM tumor
GBM1 tumor DNA vs public GBM tumor
MBCSF1 cfDNA vs public MB tumor
MBCSF2 cfDNA vs public MB tumor

A



B

Insertion location by LiBis	Support reads number	HPV type	Related Gene
Chr10:103630582-103630583	3	HPV31	SH3PXD2A
Chr16:74113811-74113812	2	HPV18	**LOC101928035**
Chr2:41807312-41807313	11		
Chr2:41905422-41905423	10	HPV18	LOC388942
Chr2:41905433-41905434	30		
Chr2:209406047-209406048	2	HPV31	**MAP2**
Chr20:10647272-10647300	2	HPV16	**JAG1**
Chr8:127732600-127732601	5	HPV16	**CASC11**

C