Published online 16 August 2019

Nucleic Acids Research, 2019, Vol. 47, No. 19 ***e122***
doi: 10.1093/nar/gkz717

DNA Analysis by Restriction Enzyme (DARE) enables concurrent genomic and epigenomic characterization of single cells

Ramya Viswanathan[1,2,†], Elsie Cheruba[1,2,†] and Lih Feng Cheow[1,2,*]

[1]Department of Biomedical Engineering, National University of Singapore, Singapore 117583, Singapore and [2]Institute for Health Innovation and Technology (iHealthtech), National University of Singapore, Singapore 117583, Singapore

Received March 26, 2019; Revised June 21, 2019; Editorial Decision August 05, 2019; Accepted August 13, 2019

ABSTRACT

Genome-wide profiling of copy number alterations and DNA methylation in single cells could enable detailed investigation into the genomic and epigenomic heterogeneity of complex cell populations. However, current methods to do this require complex sample processing and cleanup steps, lack consistency, or are biased in their genomic representation. Here, we describe a novel single-tube enzymatic method, DNA Analysis by Restriction Enzyme (DARE), to perform deterministic whole genome amplification while preserving DNA methylation information. This method was evaluated on low amounts of DNA and single cells, and provides accurate copy number aberration calling and representative DNA methylation measurement across the whole genome. Single-cell DARE is an attractive and scalable approach for concurrent genomic and epigenomic characterization of cells in a heterogeneous population.

INTRODUCTION

Genetic and epigenetic aberrations of the genome are hallmarks of cancer. Large scale copy number aberrations (CNA), resulting in structural variation of DNA sequence in somatic cells, could alter gene dosage and play critical roles in oncogenesis (1). Beyond CNA, aberrant DNA methylation in CpG dinucleotides have also been associated with suppression of tumor suppressor genes and activation of oncogenes (2,3). As such, a genome-wide characterization of genomic and epigenomic heterogeneity in single cells of disaggregated tumors could provide important insights into the biological aspects of tumor development.

Advances in sequencing technologies have shifted CNA determination from array-based technologies to more quantitative sequencing-based approaches, by which a number of single-cell CNA assays have been demonstrated. Most of the approaches to generate whole genome libraries from single cells such as Multiple Displacement Amplification (MDA), Degenerate Oligonucleotide Primed PCR (DOP-PCR) and Multiple Annealing and Looping Based Amplification Cycles (MALBAC) rely on random priming (4–7). On the other hand, Ligation-Mediated PCR (LM-PCR) Whole genome Amplification (WGA) method relies on specific restriction enzyme cutting and controlled PCR to perform deterministic amplification of the whole genome, and it was demonstrated to have improved reproducibility and reduced allelic bias (8). Utilizing this approach, a single-tube streamlined method has been developed to detect CNAs in single cells at high accuracy and resolution (9). To date, however, none of these methods are capable of concurrently measuring the epigenetic modifications encoded in the original DNA, as the information is irrecoverably lost during amplification process.

The most widely studied epigenetic modification is DNA methylation, which is commonly detected through methods that rely on bisulfite conversion or Methylation Sensitive Restriction Enzymes (MSRE). Bisulfite treatment converts unmethylated cytosines to uracil and can provide methylation information at base-pair resolution upon sequencing. Recently, single-cell Whole-Genome Bisulfite Sequencing (scWGBS) (10), genome-wide single-cell Bisulfite Sequencing (scBS) (11) and single-cell Reduced Representation Bisulfite Sequencing (scRRBS) (12) were reported to enable DNA methylation analysis in single cells. While scWGBS enables uniform coverage across the genome and could allow inference of CNAs (10), there is an associated high cost to cover a significant fraction of the ∼28 million CpG sites distributed across the human genome. scRRBS decreases the cost by restricting DNA methylation profiling to CpG-dense regions such as gene promoters and CpG islands (CGI), however it grossly underrepresents distal regu-

*To whom correspondence should be addressed. Tel: +65 6516 4158; Email: bieclf@nus.edu.sg
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

latory elements that are often more informative than canonical promoters (13). Bisulfite-based methods also have inherent limitations for single-cell assays, as the harsh chemical processing and multiple sample cleanup steps required for bisulfite-based sequencing could lead to non-specific sample loss. This is further exacerbated by poor mappability of bisulfite converted DNA sequences. Therefore, there is a need to develop alternative methods for low-input or single-cell assays that avoid sample loss associated with bisulfite protocols, while providing representative genome-wide methylation and CNA information.

An alternate method to detect DNA methylation involves the use of MSRE that selectively digest unmethylated DNA. MSRE approaches such as HpaII-tiny fragment Enrichment by Ligation-mediated PCR sequencing (HELP-seq), Methyl-seq and Methylation Sensitive Cut Counting (MSCC) (14–16) have been used to determine unmethylated regions genome-wide. A recent method, scCGI-seq, combines MSRE digestion and MDA for selective detection of methylated CGIs at the single-cell level (17). As these methods are based on counting the sequenced fragments, this poses a challenge as true negatives at a specific location cannot be distinguished from false negatives that arise from read loss. Here, we report a single-tube enzymatic method, DNA Analysis by Restriction Enzymes (DARE), that enables quantitative analysis of both unmethylated and methylated DNA in the same sample. Information of both methylation status is captured by differential adapter tagging of DNA fragments that are sequentially digested by a pair of methylation sensitive and insensitive restriction enzymes. This produces sequencing reads appended with specific tags corresponding to the methylation state of the particular CpG site. True methylated sequences can therefore be easily distinguished from read loss events. Digital counting of methylated and unmethylated reads in each library ensures precise quantification of DNA methylation levels for both low-input and single-cell samples. Due to the representative and deterministic genome coverage, DARE enables CNA calling at 500 kb resolution. This novel DARE approach will significantly augment current techniques for concurrent single-cell CNA and DNA methylation analysis.

MATERIALS AND METHODS

Cell culture

K562 cells (ATCC® CCL-243™) were cultured in high glucose Dulbecco's modified Eagle's medium (DMEM) (Gibco) supplemented with 10% Fetal Bovine Serum (FBS) (Gibco) and 1% penicillin-streptomycin (Gibco). HepG2 cells (ATCC® HB-8065™) were cultured in low glucose (1 g/l) DMEM (Gibco), 1% Glutamax (Gibco), 1% non-essential amino acids (NEAA) (Gibco), 10% FBS (Gibco) and 1% penicillin–streptomycin (Gibco). H1 ES cells was a gift from John Chua's lab. The cells were harvested, and the DNA was extracted using DNeasy blood and tissue kit (QIAGEN) for the low-input assays. DNA concentrations were quantified using Qubit dsDNA HS Assay Kit (Thermo Scientific) in Qubit 3.0 (Invitrogen). The cells were stained with CellTrace Calcein Red-Orange AM (Life Technologies), diluted in Phosphate Buffered Saline (PBS) (Gibco) to 1 cell/μl concentration and isolated in 0.2 ml PCR tubes for the single-cell assay. The tubes were observed under the microscope to confirm the presence of a single fluorescently stained cell.

Adapters design and preparation

List of adapter and primer sequences are found in Supplementary Table S1. The adapters are designed such that U-tag (unmethylated) or M-tag (methylated) adapters will be ligated on one end of the digested fragment while N-tag (NlaIII) adapter will be ligated on the other end. NlaIII digestion is included to reduce the average library size to ensure compatibility with Illumina sequencing systems. All three adapters were obtained by annealing a long and short oligonucleotide together. The long oligonucleotide of U-tag adapter and M-tag adapter consist of an 8-base Unique Molecular Identifier (UMI) followed by the respective adapter sequences to distinguish them apart in 5′ to 3′ direction, while the complementary short oligonucleotides contain 5′CG overhang. Similarly, N-tag adapter oligonucleotides consist of a long oligonucleotide with 3′CATG overhang and a short oligonucleotide. The 3′ end of the long oligonucleotides would ligate to 5′ phosphate group of the digested genomic DNA. The short oligonucleotides with low melting temperature would detach and allow extension of the digested fragment by polymerase during the first step of amplification (72°C for 13 min), completing the adapter complementary sequence. All the short oligonucleotides of the adapters consist of few uracil bases which were excised by Thermolabile USER® II enzyme (New England Biolabs), leaving the excess adapters as single stranded long oligonucleotides that are not capable of ligating. Single stranded adapter oligonucleotides were eventually removed using Exonuclease I (Enzymatics). Oligonucleotides were ordered from Integrated DNA Technologies (IDT), and the respective long and short oligonucleotides were annealed in 1× CutSmart buffer at final concentration of 25 μM (ramp down at 1°C per min from 75 to 25°C).

Sample preparation for DARE

Unmethylated Lambda DNA (Promega) was used as the unmethylated control. Methylated control was obtained by methylating unmethylated Lambda DNA (Promega) twice using CpG Methyltransferase (M.SssI) (NEB) at 37°C for 3 h. For low-input experiments, 1 μl containing 100 or 1 ng of DNA was used. For single cell experiments, cells were isolated in 1 μl of PBS into 0.2 ml PCR tubes and verified to be single cells using a microscope.

DARE workflow

1 μl of input sample containing DNA or a single cell was added into individual tubes. Details of reaction mixture compositions are found in Supplementary Table S2. All the single-cell reaction mixtures additionally contained C1 Loading Reagent (Fluidigm). 2 μl of lysis reaction mixture

with Protease (QIAGEN) was added to the tubes and incubated at 42°C for 3 h, 65°C for 30 min and 80°C for 15 min. 2 μl of HpaII digestion reaction mixture containing HpaII (New England Biolabs) was added to the tube and incubated at 37°C for 3 h to digest the unmethylated CCGG sites, and the enzyme was inactivated at 65°C for 20 min. Following digestion, 2 μl of U-tag adapter ligation reaction mixture was added and the adapters were ligated at 25°C for 2 h. This was followed by ligase inactivation at 65°C for 20 min. The excess uracil adapters were removed with 2 μl of Thermolabile USER® reaction mixture at 37°C for 20 min, 25°C for 20 min and the enzyme was inactivated at 65°C for 20 min. The remaining CCGG sites and CATG sites were then digested with 21 μl of MspI (New England Biolabs) and NlaIII (New England Biolabs) digestion reaction mixture at 37°C for 3 h, and the enzymes were inactivated at 65°C for 20 min. MspI and NlaIII digested sites were then ligated with 7 μl of M-tag/N-tag adapter ligation reaction mixture at 25°C for 2 h and T4 DNA ligase HC (Thermo Scientific) was inactivated at 65°C for 20 min. 1 μl of Thermolabile USER® II enzyme (New England Biolabs) was used to remove excess M-tag adapters and N-tag adapters at 37°C for 20 min, 25°C for 20 min, and inactivated at 65°C for 20 min. Single stranded excess adapter oligonucleotides were eliminated by the addition of 1 μl of Exonuclease I (Enzymatics) at 37°C for 30 min. The single strand ligated products were first extended with Preamplification reaction mixture at 72°C for 13 min, followed by inactivation of Exonuclease I at 85°C for 20 min. In this assay, the inhibition of Exonuclease I at 85°C was performed after the extension step. This is to prevent the fragments from denaturing before obtaining the double stranded fragments with complete tags on both ends of the fragment. Since our method relies on single stranded ligation, denaturation of DNA before extension will result in only one end of each fragment being tagged, with either U-tag/M-tag or N-tag. Next, 2 μl of 20 μM Read 1 primer and N-tag long oligonucleotide were spiked-in to the tube and amplified at 98°C 45 s, 15 cycles of 98°C 15 s, 67°C 30 s, 72°C 15 s and then 72°C 30 s. The amplified products were cleaned up with 1.6× AMPure XP beads (Beckman) and eluted in 12 μl of water. Read 1 and Read 2 primers were added to the amplicons through qPCR amplification in Reamplification reaction mixture at 98°C 45 s, X cycles of 98°C 15 s, 67°C 30 s, 72°C 15 s. For this reaction, 1 μl of 20-fold diluted product from 100 ng input sample, and 1 μl of 1 ng and single cell products was used. The number of cycles required was optimized using real-time monitoring, with 100 and 1 ng samples requiring six cycles and single cell samples requiring nine cycles. The reamplified library was purified using MinElute PCR Purification Kit (QIAGEN). Qubit dsDNA HS Assay Kit (Thermo Scientific) was used to quantify the library concentration and equal concentration of all libraries were pooled. Products between 180bp-420bp were size selected with BluePippin 2% agarose cassette (Sage Sciences) and purified using MinElute PCR Purification Kit. Qubit dsDNA HS Assay Kit was used to quantify the library concentration. The size distribution of the library was measured using Bioanalyzer High Sensitivity DNA Analysis Kit (Agilent). This was followed by quantification of the library with Kapa Library Quantification Kit (Roche) and sequenced in MiSeq or HiSeq 4000 sequencing system (Illumina).

Data processing

To process the sequencing data to obtain methylation values, we developed a pipeline consisting of the following main steps: (i) extracting the UMI, (ii) differentiating methylated and unmethylated reads based on the adapter sequence, (iii) adapter trimming, (iv) alignment to the human genome, (v) removal of PCR duplicates, (vi) calculating methylation ratio at each CCGG site. The quality of the sequenced reads was analyzed using FastQC v0.11.4 (18). UMI-tools v0.5.4 was used to extract the UMI and remove PCR duplicates (19). Adapters were trimmed using Cutadapt v1.5 (20) and alignment to hg38/GRCh38 human reference genome was done using bowtie2 v2.3.4.1 (21).

Ploidy determination and copy number aberration analysis

The human genome was split into 500 kb windows. This resolution was chosen as 93.5% of the 500 kb windows were covered by 50 unique fragments or more. The sum of unique reads in each window was divided by the number of assayable sites in that window to obtain the read counts by window. The ploidy determination approach was adapted from Kendall and Krasnitz (22). Using the reads count by window, a range of multiplier values were chosen such that $1.5 \leq \langle MS \rangle \leq 5.5$, where M is the multiplier and S denotes the reads counts. Subsequently, the mean squared rounding error $\langle (MS - [MS])^2 \rangle$ was calculated for each multiplier and the multiplier at which this error was minimum was chosen (Supplementary Figure S1A). The estimated copy number for each bin was calculated using the previously obtained multiplier and the median ploidy was given as the highest density integer (Supplementary Figure S1B). In order to validate our approach, the ploidy determination was performed on H1 ES cells, which yielded a consistent copy number of 2 (Supplementary Figure S1C).

To obtain the final copy number values for each 500 kb window, the reads for each cell were normalized using the median of that cell and multiplied by the obtained ploidy. To provide integer copy number for use in Circos plot, the obtained values were rounded to the nearest integer.

CNA calling accuracy

To determine the accuracy of the CNA calling algorithm, reference HepG2 CNA values from Zhou *et al.* were used (23). The reference CNA values were binarized as normal and aberrated for this purpose. Our determined CNA values for each window was binarized using different thresholds and the sensitivity and specificity was calculated at each of these thresholds. Using the obtained values, a Receiver Operating Characteristics (ROC) curve was plotted and area under the curve (AUC) was calculated to determine accuracy.

Single nucleotide variation analysis

Whole genome sequencing data was obtained from ENCODE (ENCFF336CFC). The SNV calling was done us-

ing samtools mpileup (24). Our sequenced library was compared with hg38 reference to determine the variations.

Determining DNA methylation

A given CCGG site was considered assayable if it had a CATG site more than 32 bp and less than 272 bp away from the CCGG site of interest, and no other CCGG site preceding the CATG site.

The DNA methylation at each CCGG site was calculated as the unique number of reads with M-tag divided by the total unique reads at that site.

$$Methylation\ ratio = \frac{\#\ M-tag}{\#\ (U-tag+\ M-tag)}$$

Confirming accuracy of DNA methylation

Whole Genome Bisulfite Sequencing (WGBS) data for K562 and HepG2 was downloaded from ENCODE (K562: GSM2308596_ENCFF721JMB, HepG2: ENCFF369YQW). WGBS samples were filtered to include only sites with five or more reads to ensure reliable methylation values. For the merged sample generated by combining the single cells, only sites with five or more covered reads were included in the analysis. Different gene region annotations, such as CpG Islands, introns, exons and repeats were downloaded from UCSC table browser. Promoters were defined as the region that is 1 kb downstream and 500 bp upstream of the transcription start site of each gene (25). To calculate mean methylation for different regions, the average methylation of the CCGG sites covered within the region were used. For scDARE, pairwise Pearson correlation for complete observations was calculated between all the single cells.

To visualize the Pearson correlation between DARE and other assays, we employed density scatter plots. The CpG sites are binned by methylation percentage and the color scale represents the densities of CpG sites of particular methylation percentages. This approach avoids overplotting of similar valued CpG sites that obscure the structure in the bimodal distributed DNA methylation data. The Pearson correlation score is generated from the raw data.

Comparison with other methylation assays

RRBS data was downloaded from ENCODE (ENCFF001TNA). Other single-cell methylation assay data was obtained from the following NCBI GEOs (scRRBS: GSE47343, scBS: GSE56879).

Gene body methylation analysis

To correlate the gene body methylation with expression of the respective genes, HepG2 RNA-Seq data was obtained from ENCODE (ENCFF004HYK). Genes were ranked by 'transcripts per million (tpm)'. Genes with less than 1 tpm were denoted as 'low expression' while genes with >100 tpm were denoted as 'high expression'. All other genes with tpm between 1 and 100 were denoted as 'intermediate expression'.

Allele-specific methylation

SNVs from DARE methylated and unmethylated reads were identified using samtools mpileup and filtered for significance using vcfutils. Only sites with heterozygous SNVs was considered. Fisher's exact test and q-value <0.05 was used to select regions with allele-specific methylation.

Saturation plot

The cumulative CCGG coverage of DARE was determined by a saturation plot. Starting from a single randomly chosen sample, the number of CCGG sites covered was calculated. Additional experiments were randomly added and the cumulative number of unique CCGGs covered was calculated at each step. This process was repeated ten times and the results obtained was plotted.

RESULTS

DARE workflow and adapter design

DARE is a LM-PCR based approach to methylation-specific WGA. The basic principle of DARE is the methylation-specific tagging of CCGG sites across the genome as illustrated in Figure 1A. This is achieved by first digesting genomic DNA with methylation-sensitive HpaII enzyme that cleaves unmethylated CCGG sites. The cleaved fragments are ligated with a uracil-containing double-stranded 'U-tag' adapter. Subsequent treatment with Thermolabile USER® II enzyme inactivates the unligated U-tags by converting them into single-stranded oligonucleotides. Following this, genomic DNA is digested with methylation-insensitive MspI enzyme that digests the methylated CCGG sites and NlaIII enzyme that further fragments the genomic DNA by cutting at CATG sites. MspI and NlaIII cut sites are ligated with 'M-tag' and 'N-tag' adapter respectively, both of which are uracil-containing double-stranded adapters. All unligated adapters are made single-stranded and removed from solution by treatment with Thermolabile USER® II enzyme and Exonuclease I. At this point, all the genomic DNA molecules are ligated with N-tag adapter at CATG sites, U-tag adapter at unmethylated CCGG sites and M-tag adapter at methylated CCGG sites. Accordingly, PCR amplification of these tagged molecules produces three types of products: (i) U-tag/M-tag adapter on one end and N-tag adapter on the other, (ii) U-tag/M-tag adapter on both ends, (iii) N-tag adapter on both ends. Of these products, only the first type is efficiently amplified and sequenced, while the rest are inefficiently amplified due to PCR suppression effect. This deterministic amplification of fragments is a unique feature that is important for consistent methylation comparison and CNA determination across different samples.

The features of the adapters are shown in Figure 1B. The U-tag adapter and M-tag adapter consists of several important regions: (i) Read 1 primer sequence, (ii) 8-base Unique Molecular Identifier (UMI), (iii) Tag region (TTAGCGAC ACGA for U-tag, AGCAGATGACGT for M-tag) and (iv) 5′CG overhang for specific ligation to HpaII/MspI cut sites. Meanwhile, the N-tag adapter consists of Read 2 primer

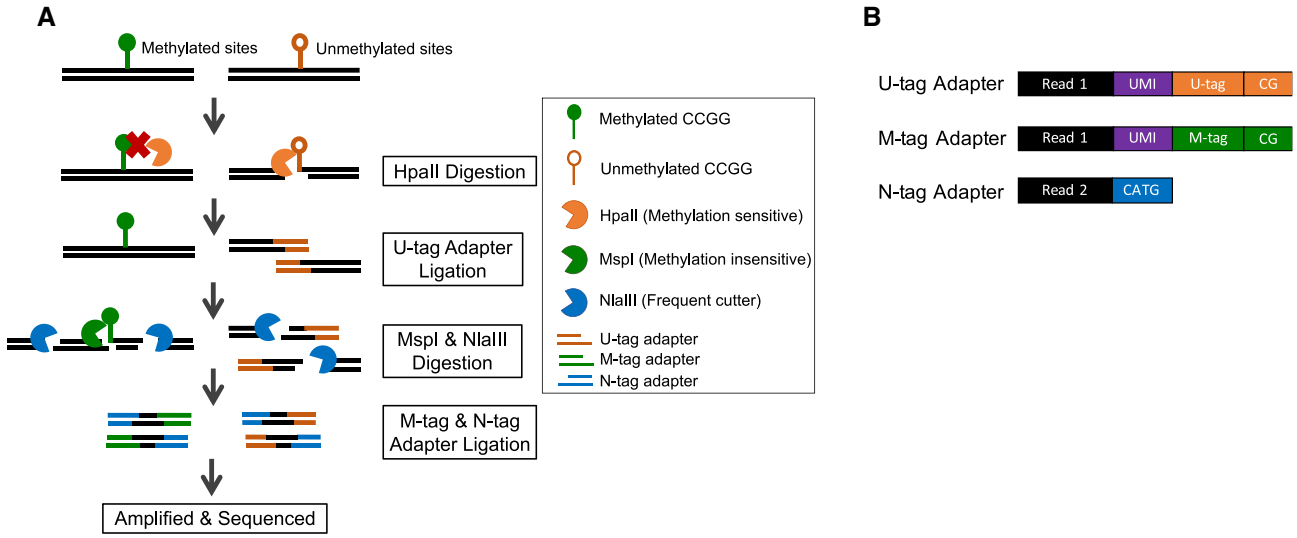

Figure 1. Workflow of DNA Analysis by Restriction Enzyme (DARE) assay. (**A**) Workflow of DARE assay—cell lysis and protease treatment are followed by digestion of unmethylated CCGG sites with methylation sensitive HpaII enzyme. U-tag adapters are ligated and the remaining CCGG sites are digested by methylation insensitive MspI enzyme. NlaIII digestion is included to reduce the fragment length. This is followed by ligation with the respective adapters (M-tag and N-tag adapters). Thermolabile USER® II enzyme is used to remove excess uracil-containing adapters after each ligation. (**B**) Adapter system: U-tag adapter consists of Read 1 primer sequence of Illumina adapter, unique molecular identifier (UMI), unmethylated site specific tag (U-tag), and CG overhang. M-tag adapter similarly consists of Read 1 primer sequence of Illumina adapter, UMI, methylated site specific tag (M-tag), and CG overhang. N-tag adapter consists of Read 2 primer sequence of Illumina adapter and CATG overhang.

sequence and 3′CATG overhang for ligation to NlaIII cut sites. This particular adapter design has many advantages especially for low-input samples: (a) it enables sticky end ligations that are much more efficient than blunt-end or TA ligation, (b) the use of non-phosphorylated adapters eliminates adapter-adapter ligation and (c) the UMIs reduce the effect of quantitative biases in low-input sample.

To assess the feasibility of DARE assay to distinguish methylated from unmethylated CCGG sequences, we performed the assay on unmethylated and enzymatically-methylated lambda DNA. As expected, the majority of reads obtained from the unmethylated control samples contain U-tag, while the majority of reads obtained from the methylated control samples contain M-tag (Supplementary Figure S2A). As a control of enzyme digestions in our reaction tubes, we spiked in unmethylated lambda DNA for estimation of enzyme efficiency.

Accurate copy number aberration determination using DARE

To benchmark this novel assay on low amounts of DNA, we first performed DARE on 100 and 1 ng of HepG2 genomic DNA according to the protocol described (Supplementary Table S2). On average, 12.3 million reads per cell (8.3–16.1 million, Supplementary Table S3) were obtained and mapped to hg38, with an average mapping efficiency of 93% (Supplementary Table S3). We first assessed the genome-wide distribution of DARE fragments for read-count based CNA calling. *In-silico* digestion of the human genome yielded 1 806 438 fragments in the 32–272 bp range that are flanked by CATG on one side and CCGG on the other. Due to the deterministic amplification of DARE fragments, we performed read-count normalization based on *in-silico* fragments for CNA calling. As the signal-to-noise ratio of CNA calling is related to the number of reads per window, we determined the fraction of the genome that is covered by >50 *in silico* DARE fragments for different window size (Figure 2A). At a resolution of 500 kb, 93.5% of the windows across the genome satisfy this criterion and is expected to provide robust CNA calling. The same trend is observed in the 100ng HepG2 DARE sample as the *in-silico* simulation. We also used median absolute pairwise difference (MAPD), which indicates the noise in CNA calling at a particular resolution. At the resolution of 500 kb, MAPD for 100 ng HepG2 DARE sample was calculated to be 0.10, which is within the required 0.45 cutoff for CNA calling (26). As a control, we performed DARE assay on 100 ng gDNA from diploid H1 ES cells, subjected to the same bioinformatics analysis for CNA calling, and showed that a diploid copy number is obtained across the entire genome. The CNAs for HepG2 and H1 ESCs obtained from DARE are visualized in the Circos plot in Figure 2B.

To verify the accuracy of CNA calling, we compared the copy number obtained with DARE at 500 kb resolution against published copy number data for HepG2 (23). Comparison of corresponding copy number values from DARE and reference data demonstrated a high degree of concordance (Figure 2C). ROC analysis showed an excellent agreement (AUC = 0.96) between CNA calls from DARE and reference dataset (Figure 2D).

Assaying for DNA methylation in low-input samples using DARE

The key feature of the DARE assay is the retention of DNA methylation information even upon WGA. It offers some advantages over other methylation assays: Firstly, the use of U-tag and M-tag to identify methylation status of each

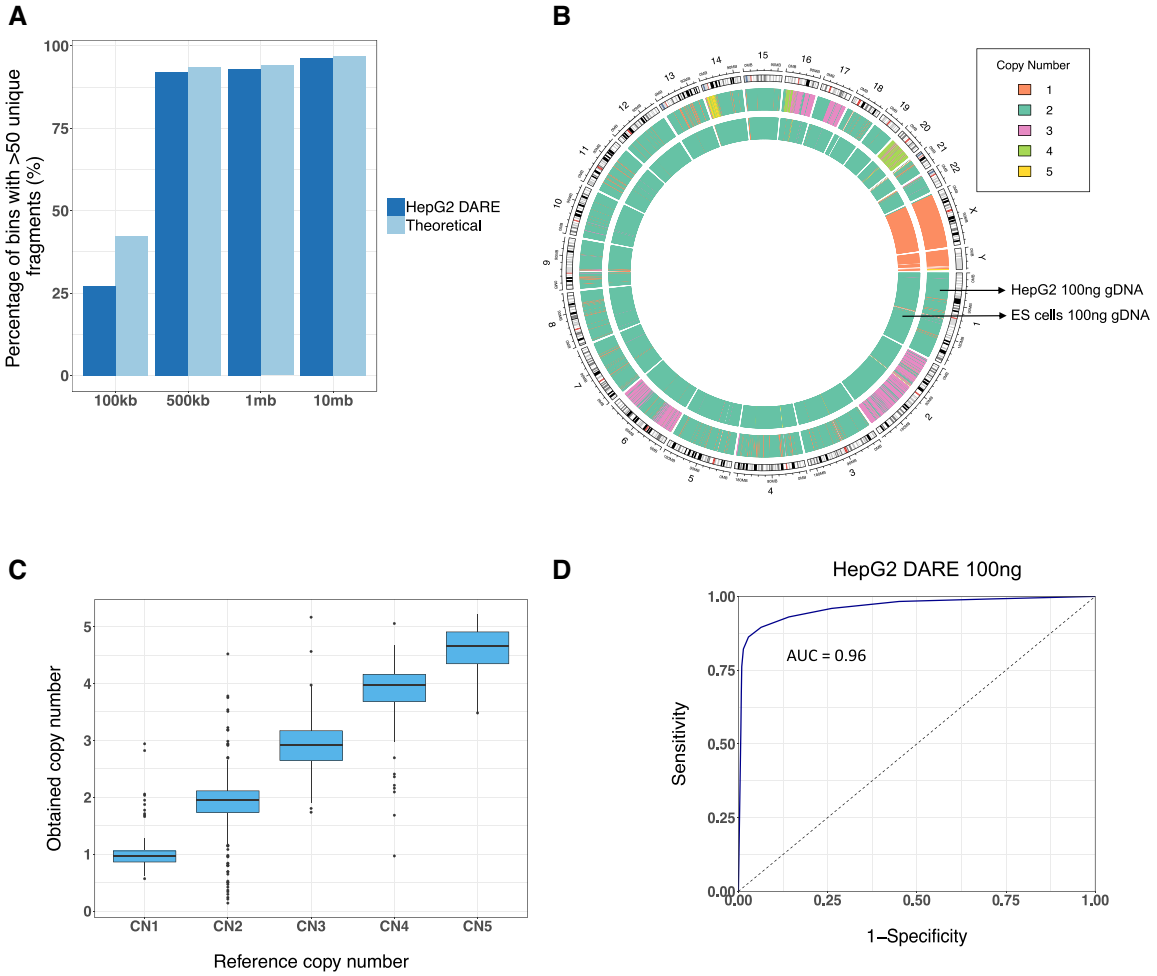

Figure 2. Unbiased read distribution of DARE and its accuracy in determining copy number aberrations. (**A**) Percentage of 100 kb, 500 kb, 1 mb and 10 mb windows that contain 50 or more theoretical assayable sites and 50 or more unique fragments per bin in the 100 ng HepG2 DARE sample. (**B**) Circos plot showing CNAs in DARE 100 ng HepG2 and H1 ES cells gDNA samples. The outermost circle is the cytoband of the human genome. The middle circle is the copy number of HepG2 at 500 kb resolution and inner circle is the copy number of H1 ES cells at 500 kb resolution. (**C**) Box plot of obtained copy number from DARE HepG2 100 ng sample and corresponding values in the reference data. (**D**) ROC analysis of obtained CNA values at 500 kb resolution from DARE HepG2 100 ng DNA, AUC = 0.96.

CCGG site allows for direct determination of the methylation status without requiring normalization based on a separate sample like most MSRE-based methods (HELP, MSCC, scCGI-seq etc.) (Supplementary Figure S2B). Secondly, the high complexity of insert sequences results in high mapping efficiency (93%) compared to the bisulfite assays (55–73%) (27,28).

The MSRE-based DARE assay can only profile the methylation states of CpG sites located in the CCGG context. Theoretical analysis showed that while there are ∼28.7 million CpG sites in the human genome, ∼2.3 million (8%) of these are located in CCGG context. After filtering for fragment size and removing sequences that map to multiple locations, 1 385 655 unique CCGG sites are theoretically covered by DARE. This represents 4.8% of the ∼28.7 million total CpG sites or 59.7% of the ∼2.3 million CCGG sites in the human reference genome. Thus, while DARE covers far less CpG sites than WGBS, its theoretical CpG coverage is comparable to other methylation detection technologies such as HumanMethylation 450 BeadChip (∼450 000 CpGs) and RRBS (∼3.4 million CpGs). A comparison between CpG coverage of DARE and other technologies is summarized in Supplementary Table S4.

DARE assay on 100 ng HepG2 DNA, sequenced to 11.4 million reads, covers 86% (1.19 million/1.38 million) of the assayable DARE CpGs, with average sequencing depth of 6. The genomic distribution of assayable DARE CpGs closely tracks the total CpG distribution, and provide an excellent representation of the distribution of CpGs in different genomic elements. To illustrate the linear spatial relationship between CCGG and CpG sites, we showed that the number of CpG sites between consecutive CCGG sites is tightly distributed (Supplementary Figure S2C), and the frequencies of CpG sites and CCGG sites within 10 kb windows are highly correlated (Supplementary Figure S2D). The analysis of a single CpG site or a few CpG sites has been widely used as surrogate indicators of the DNA methylation status of the corresponding element (29). The correlation between the average DARE CCGG methylation and average WGBS CG methylation at CGIs was 0.84, indicating a high

predictive sensitivity of this approach. We further showed that the genomic representation of DARE libraries closely matches that of WGBS and recapitulates the theoretical distribution of different genomic regions (Figure 3A). Compared to RRBS where ∼80% of promoters are represented by one or more measurements (30), DARE achieves similar coverage where 74% of the promoters are represented by one or more DARE CpG (Supplementary Figure S2E).

To validate the accuracy of the methylation values obtained by DARE, we compared the values to that of WGBS data. The Pearson correlation of methylation values at individual CCGG sites between HepG2 WGBS and DARE assay from 100 ng HepG2 DNA was 0.92 (Figure 3B), while correlation between replicate 100 ng HepG2 DNA samples was 0.93 (Supplementary Figure S2F), indicating good accuracy and reproducibility. The accuracy was maintained even with a lower amount of input DNA, with the correlation between 100 and 1 ng HepG2 DNA being 0.92 (Supplementary Figure S2G), showing that DARE can be applied to small amounts of input DNA for accurate epigenome profiling.

We summarized DARE DNA methylation values across gene bodies and 15 kb regions upstream and downstream of each gene and detected the characteristic hypomethylation valleys around transcription start sites (TSSs) as well as hypermethylation patterns of the gene bodies (Figure 3C). Beyond promoters, there has recently been many studies to establish the effect of gene body methylation on gene expression (31–33). By analyzing HepG2 RNA-Seq data from ENCODE (ENCFF004HYK) and methylation information obtained through DARE, we observed that lower expressing genes had intermediate methylation level of gene body, while the intermediate and high expressing genes had high gene body methylation (Figure 3D and E). This was also observed in the WGBS data, validating the accuracy of the trend observed (Supplementary Figure S2H and I). As DARE provides proportional representation across the genome, it serves as a valuable tool to investigate the function of DNA methylation in different genomic contexts.

Concurrent genomic and epigenomic analysis detects allele specific methylation

Apart from CNA determination, the excellent mappability of non-bisulfite-treated DARE fragments enables robust SNV calling for simultaneous genomic and epigenomic analysis. In regions covered by ≥10 DARE fragments, we detected 89% (81 099/90 894) of SNV calls made by whole genome deep sequencing of HepG2 cells DNA, indicating the high sensitivity of DARE genomic analysis. Of these SNVs, 35% (29 110/81 099) were C→T or G→A conversions that would otherwise be difficult to distinguish from induced deamination of cytosine in bisulfite approaches. We also investigated the frequency of SNVs in the CCGG context that might affect the performance of DARE assay. Using the HepG2 cell line as an example, we found that 30 594 of the ∼4.2 million annotated SNVs overlapped with the theoretical 1.38 million assayable CCGG sites in DARE. Therefore, only a negligible percentage (∼2.2%) of the assayable sites are potentially affected.

Recent studies have found sequence-dependent CpG methylation differences at heterozygous regulatory sequences that could lead to complex traits in human populations. However, directly identifying SNVs from bisulfite converted sequence reads is challenging and it was estimated that an average 30× sequence read depth would be required to call SNVs accurately from bisulfite sequencing data (34). This problem is circumvented in the bisulfite-free DARE assay. More than half (45 867/81 099) of the SNVs detected in DARE assay for HepG2 DNA are heterozygous and could serve as allelic markers. We performed allele-specific methylation analysis on these heterozygous SNV loci and estimated significance by means of Fisher's exact test on the counts of methylated and unmethylated cytosines observed on the same sequencing read with each of the SNV allele. Sixty eight loci with allele-specific methylation were detected with q-values of <0.05. The locations of these loci are visualized in Figure 3F and a complete list of these loci is listed in Supplementary Table S5.

As a new finding of this analysis, we observed a very striking colocation of the allele-specific methylation loci and genomic regions harboring large-scale CNA in HepG2. While only 23.5% of the HepG2 autosomes are measured by DARE to harbor CNAs at 500kb resolution, 64.7% (44/68) of the allele-specific methylation loci are found in the CNA regions (Supplementary Table S5), representing a significant enrichment. We believe that this is the first time such relationship has been reported in literature, and it is made possible by DARE assay that concurrently measures genomic CNAs, SNVs and DNA methylation in the same sample.

Application of DARE in single cells to determine CNA

On the basis of low-input DARE results, we next applied single-cell DARE (scDARE) to 10 HepG2 and 10 K562 single cells. On average, we obtained 11.1 million reads per cell with an average mapping efficiency of 89% (Supplementary Table S3), significantly higher than bisulfite-based single-cell methylation assays (24–32%) (27,35). First, we investigated the performance of scDARE for single-cell CNA detection. Between two cells of the same cell line, there was a high degree of correlation between the normalized read counts at 500 kb window resolution (Figure 4A and B), indicating a high degree of reproducibility. As a measure of data quality, we calculated the MAPD scores for scDARE samples to be 0.22 ± 0.03. Low MAPD values (<0.45) indicate low noise and evenness of whole genome amplification of this method. Following the similar approach as earlier, the median ploidy for the HepG2 and K562 cells was calculated to be 2 and 3 respectively (Supplementary Figure S3A and B), in agreement with published literature (23,36). ROC analysis showed good agreement between CNA calls from scDARE and reference dataset (AUC $= 0.87 \pm 0.05$) (Supplementary Figure S3C) as well as between scDARE and DARE (AUC $= 0.86 \pm 0.04$) (Supplementary Figure S3D). The CNA values obtained were visualized by chromosome and different regions of amplification and deletion were observed in single cells from each cell line in agreement with published literature (Figure 4C) (23,36). To estimate the sensitivity of CNA calling depending on total sequenc-

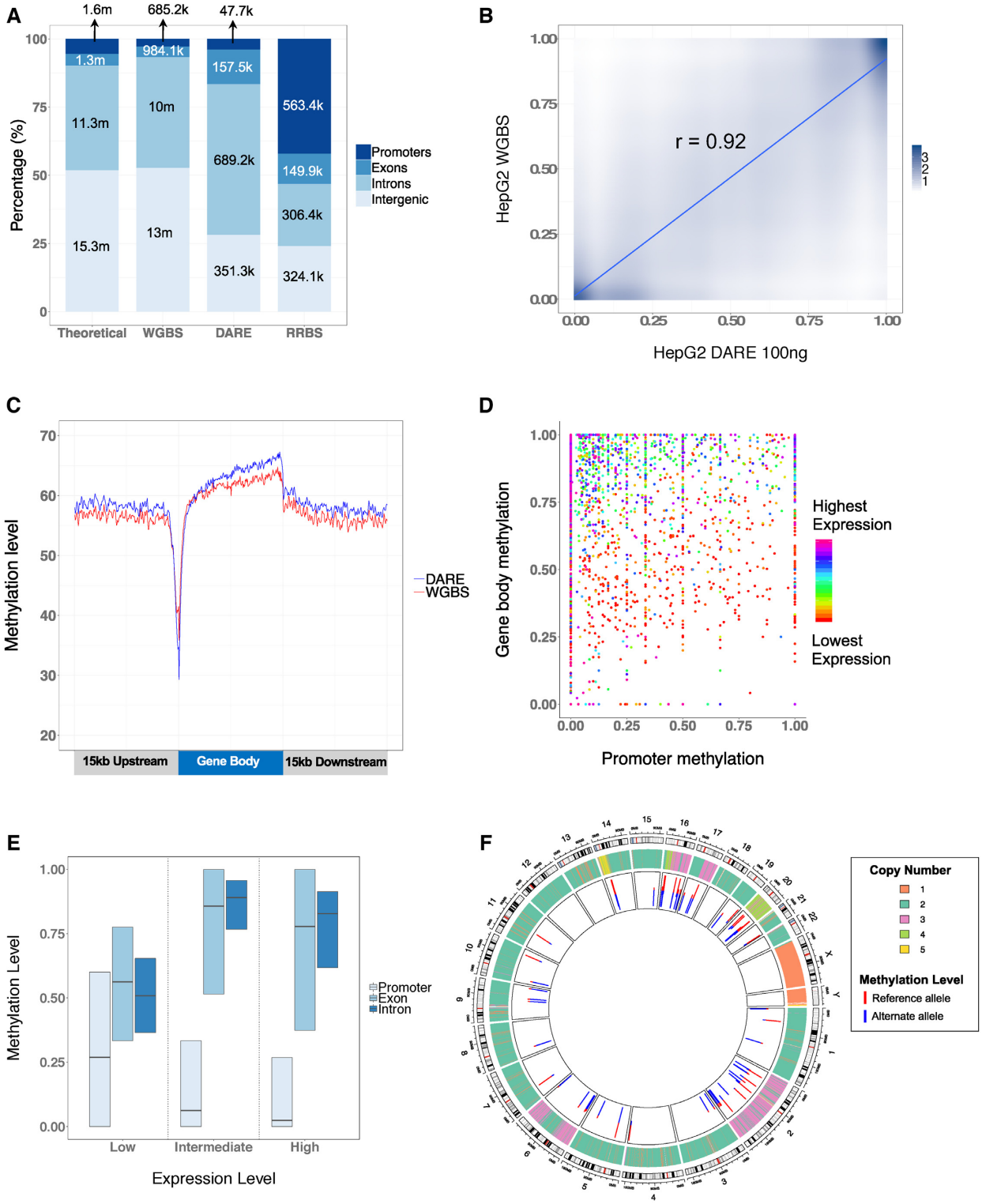

Figure 3. Unbiased coverage and DNA methylation accuracy of DARE assay from 100 ng HepG2 genomic DNA. (**A**) Theoretical distribution of promoters, exons, introns and intergenic regions in the genome, as well as in WGBS, DARE and RRBS sequenced library. (**B**) Pearson correlation coefficient of methylation values at CCGG sites in HepG2 100 ng DNA with ENCODE WGBS data. Only sites with five or more reads were considered. Pearson $r = 0.92$. (**C**) Average DNA methylation of CpGs in CCGG context of genes and its 15 kb upstream and downstream region for HepG2 DARE and ENCODE WGBS data. (**D**) Promoter and gene body methylation profile of individual genes ranked by expression level. (**E**) Range of methylation values of promoter, introns and exons, stratified by gene expression level. (**F**) Circos plot showing CNAs in DARE 100ng HepG2 (outer circle) and the loci of allele-specific methylation (inner circle). Height of the red bar represents the methylation level of the reference allele and height of the blue bar represents the methylation level of the alternate allele at these regions.

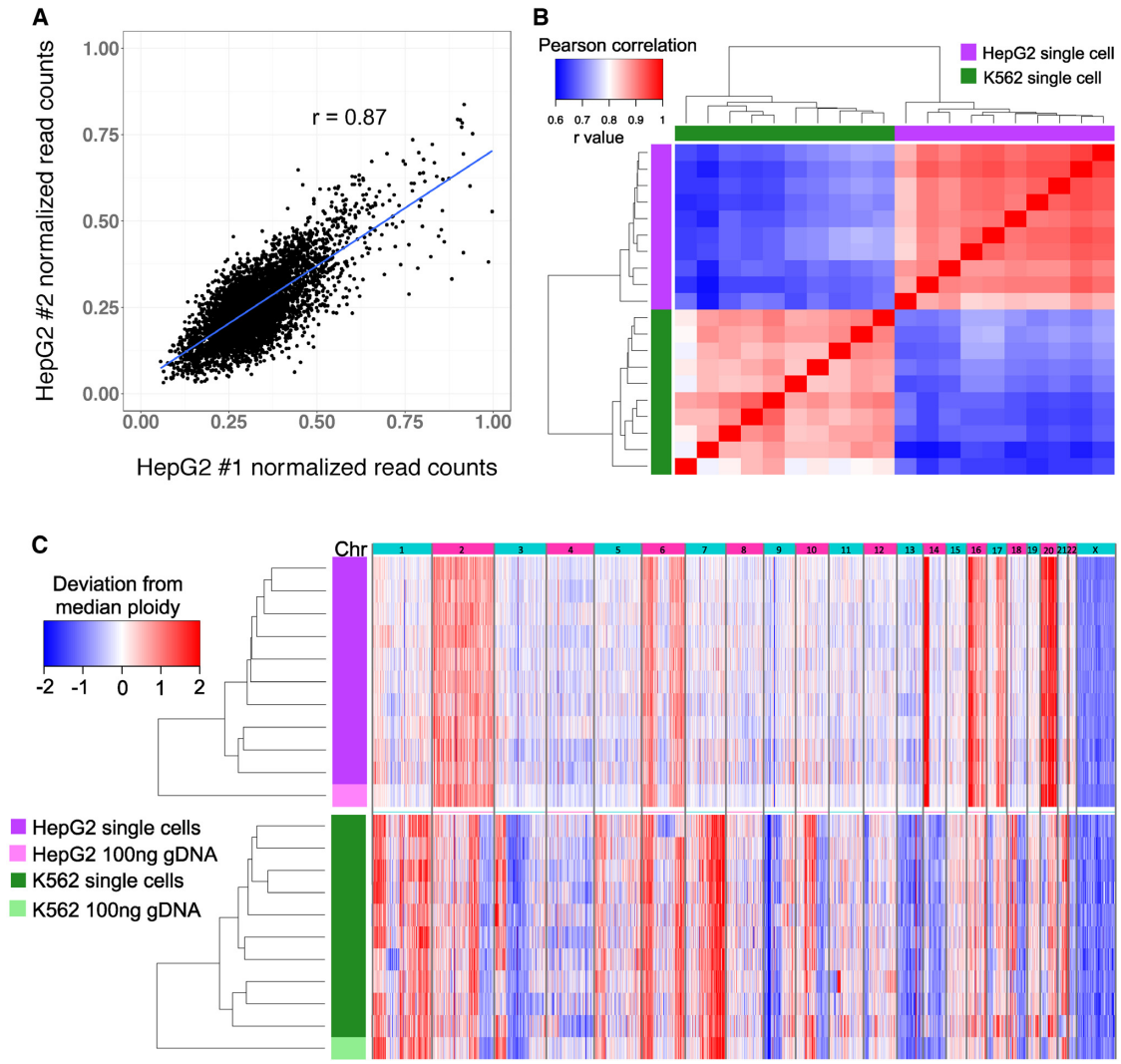

Figure 4. CNAs in single cells. (**A**) Scatter plot of normalized read counts in 500 kb for two HepG2 single cells. Pearson $r = 0.87$. (**B**) Pairwise correlation matrix of copy number for all HepG2 (purple) and K562 (green) single cells. (**C**) Heat map of CNAs detected at 500kb resolution for HepG2 (purple) and K562 (green) 100 ng gDNA and single cells. Red represents amplification and blue represents deletion with respect to its median ploidy.

ing reads, we performed CNA calling with subsets of reads ranging from 0.2 million to 10 million mapped reads extracted by random sampling of reads of a single HepG2 cell. As expected, CNA calling performance deteriorates with decreasing sequencing reads (Supplementary Figure S3E) due to reduced coverage. Nevertheless, at 1 million reads (513 318 unique reads) for a single cell, a high AUC of >0.90 is obtained, indicating that low pass sequencing for high throughput scDARE is feasible.

DNA methylation in single cells

On the methylation front, scDARE was able to cover an average of 27% (381 033/1.38 million) of assayable DARE CpG sites per cell. On merging data from 10 HepG2 single cells, we were able to cover more than 90% of the assayable sites (Supplementary Figure S4A), with an average read depth of >4 (Supplementary Figure S4B). Comparing the methylation values of CCGG sites with read depth ≥5 of the merged single HepG2 cells data with DARE results from 100 ng HepG2 DNA, we obtained a high Pearson correlation of 0.87 (Supplementary Figure S4C). We compared the global methylation of the two cell lines and observed that K562 was hypomethylated, in agreement with literature (Supplementary Figure S4D) (37). Also evident is the binary nature of the methylation profile at the single-cell level, which is observed in all the single cells assayed (Figure 5A) compared to low-input samples. We then compared the representation of different genomic regions by scDARE and other single-cell bisulfite sequencing based assays (Figure 5B) and showed that scDARE had comparable representation in most genomic regions.

DNA methylation across the genome is typically variable according to its functional context. To visualize this, we plotted the distribution of HepG2 single-cell methylation across a selected region of chromosome 1. It can be seen that methylation is homogeneously low in CGIs within promoters, and becomes homogeneously high in the gene

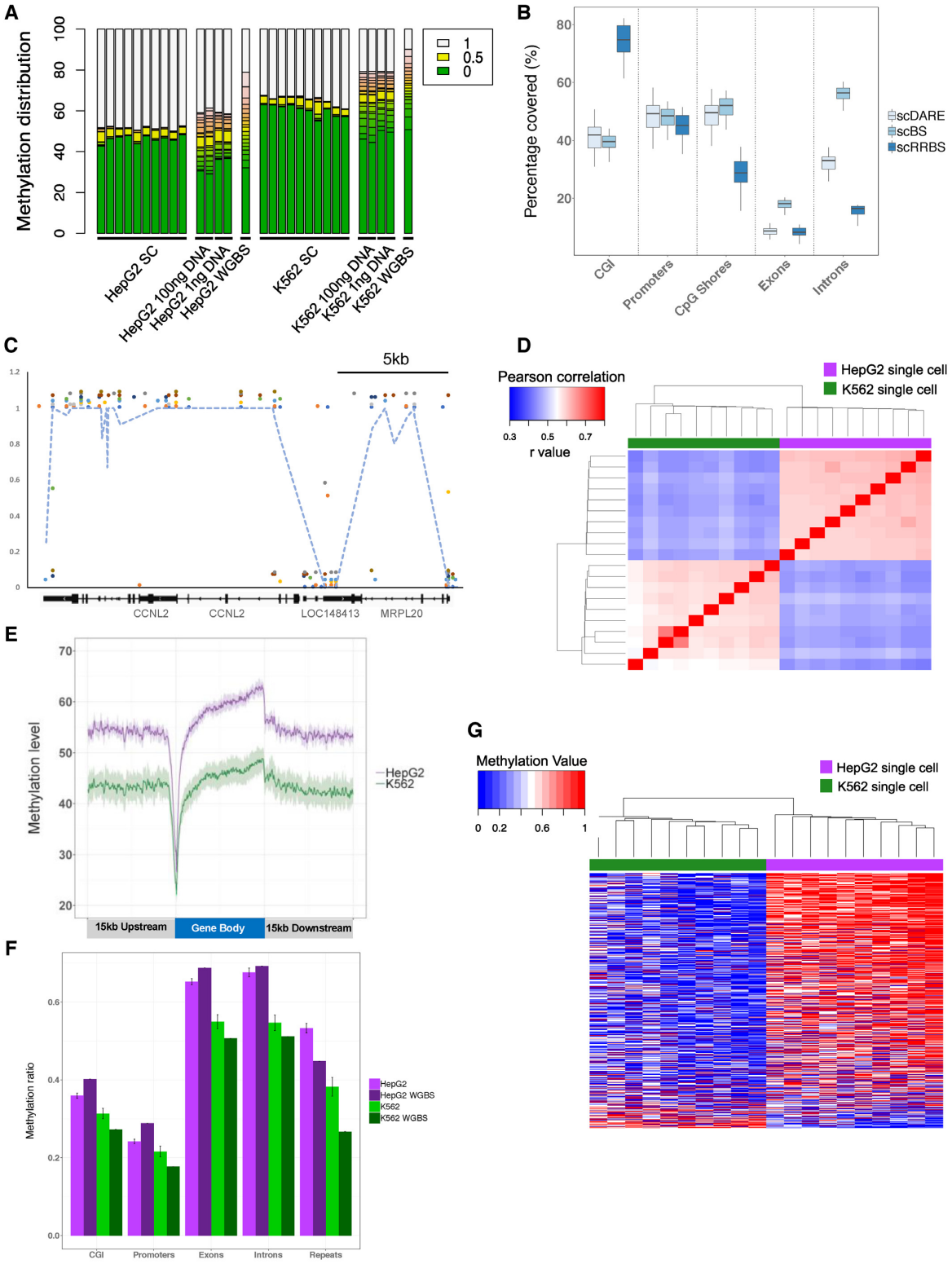

Figure 5. DNA methylation in single cells. (**A**) Distribution of DNA methylation values in scDARE, DARE low-input and WGBS samples. (**B**) Percentage of genomic regions covered by scDARE, scRRBS, scBS from 10, 8, 10 single cells respectively. (**C**) DNA methylation rates for each HepG2 cell (represented by different colour) in a short region of chromosome 1. The methylation rates of ENCODE WGBS data is shown by the blue dotted line. The region shown as an example includes the Cyclin L2 (CCNL2), and mitochondrial ribosomal protein L20 (MRPL20) locus. (**D**) Pairwise correlation matrix for all single-cell HepG2 (purple) and K562 (green) DNA methylation at assayed CCGG sites. (**E**) Average DNA methylation of CpGs in CCGG context of genes and its 15 kb upstream and downstream regions. (**F**) Average DNA methylation across functional genomic regions of 10 HepG2 and K562 single cells compared to the respective WGBS ENCODE data. Error bar represents the standard deviation. (**G**) Unsupervised clustering of the 1000 most differentially methylated 1 kb regions between K562 (green) and HepG2 (purple) covered by at least 8 cells of each cell line.

body regions (Figure 5C). Greatest cell-to-cell heterogeneity is found at CpG shores flanking the CGIs. High inter-cell type variation and low intra-cell type variation are observed between individual cells from cell lines (HepG2 and K562), based on Pearson correlation analysis of CCGG methylation levels between single cells (Figure 5D). On a global scale, individual K562 cells have significantly lower DNA methylation compared to their HepG2 counterparts, especially in gene bodies and intragenic regions (Figure 5E), consistent with trends observed in WGBS assays (Figure 5F).

Cell-to-cell comparison of single cell DNA methylation data at individual CpG level is challenging due to the sparse coverage. Aggregation of multiple measurements across genomic windows could allow for quantitative cell-to-cell analysis of DNA methylation variation. We show that a third of the 10 kb windows containing DARE CpGs are covered by 80% or more single cell samples for comparison of regional methylation levels between individual cells (Supplementary Figure S4E). We examined the 1000 most variable 10 kb regions among all the single cells, and showed that the methylation profiles of these regions robustly separate the HepG2 and K562 cells (Figure 5G). This result demonstrates that differences in cell types are encoded in their epigenetic profiles, and can be detected at the individual cell level using scDARE. Similarly, the single cells cluster according to cell types when examining the 1000 most variably methylated promoters (Supplementary Figure S4F), corroborating the role of epigenetic control that underlie gene regulation of different cells.

DISCUSSION

Here we report DARE, a novel LM-PCR Whole Genome Amplification-based approach to assay for genetic and epigenetic alterations in low-input DNA and single cells. Methylation-specific whole-genome amplification is achieved by sequential DNA digestion with methylation sensitive/insensitive isoschizomers followed by ligation of specific sequence tags. DARE provides deterministic and proportional coverage across the genome for robust read-count based copy number calling. At the same time, this is the first report of a bisulfite-free method that can simultaneously provide genome-wide DNA methylation information at base-resolution in single cells. We provide proof-of-principle of this novel multimodal assay by simultaneously detecting copy number variations and consistent DNA methylation differences between single cells from two different cell lines.

Fundamentally, DARE is an LM-PCR to whole (epi)genome amplification. The deterministic amplification properties of LM-PCR has been shown to provide superior performance compared to other WGA approaches in terms of reduced allelic bias and dropouts (8)—important considerations in single-cell assays. Judicious choice of enzymes in DARE results in excellent genome coverage: >93% of 500kb windows are covered by >50 unique DARE fragments, this enables robust CNA determination at 500kb resolution.

Nonspecific sample loss due to harsh chemical treatment and multiple sample cleanup is an inherent limitation of bisulfite sequencing, and becomes a critical bottleneck for comparison of methylation status at specific CpGs in multiple low-input samples including single cell applications. These problems are resolved in the DARE approach, due to the mild restriction enzyme steps, high efficiency of sticky-end ligation and single-tube protocol. Although DARE covers only ∼4.8% (1.8 million) of the total CpG sites, these sites are consistently covered due to the low loss nature of the protocol. Furthermore, due to the strong spatial correlation of CpG methylation states, specific methylation measured at DARE CpGs can be used to infer regional methylation. Thus, DARE directly profiles the methylation states of CCGG sites while indirectly provides information on the surrounding region.

Decrease in DNA complexity upon bisulfite conversion can lead to significantly reduced alignment rate and ambiguity in base calling. In the bisulfite-free DARE approach, consistent high mapping efficiencies are obtained in both low-input and single-cell samples. Sequencing reads from DARE can be used for genetic analysis including robust SNV calling. We detected 89% (81 099/90 894) of SNV calls made by whole genome deep sequencing in high-coverage DARE fragments on DNA from HepG2 cells. Of these SNVs, 35% (29 110/81 099) were C→T or G→A conversions that would otherwise be difficult to distinguish from induced deamination of cytosine in bisulfite approaches. Concurrent genomic (SNV) and epigenomic (CpG methylation) analysis using DARE also enabled allele-specific methylation analysis. We observed a significant enrichment of loci with allele-specific methylation at genomic regions harboring CNAs. Although further work is needed to better understand this phenomena, our hypothesis is that allele-specific methylation could play a role in silencing the aberrantly amplified genome copies, thus providing gene dosage compensation. The most well-known dosage compensation by DNA methylation mediated silencing is X chromosome inactivation in female. Our experimental observations indicate that analogous mechanisms could take place in cancer cells.

There have been several alternative approaches that make use of the HpaII/MspI isoschizomers pair to elucidate methylation information (14,16,38,39). Unlike bisulfite-based methods, these approaches often employ separate enrichment and detection of methylated and unmethylated DNA, which limits the quantitative precision of the analysis. A novel technique, DREAM, employs sequential DNA digestion with methylation sensitive SmaI and methylation insensitive XmaI to directly measure DNA methylation (40). However, the frequency of assayable sites afforded by these 6-base cutters was low, being able to cover only ∼0.15 million CpGs in the human genome. In comparison, DARE utilizes a frequent 4-base cutting HpaII/MspI pair to increase the assayable CpGs to ∼1.38 million, comparable to the ∼2.5 million CpG sites that can be assayed by RRBS (41).

Incomplete DNA digestion by HpaII could lead to occasional errors in methylation state calling in MSRE-based approaches. Using the normally unmethylated mitochondrial DNA as an internal control, we observed low percentages of <10% M-tag adapter in all except one single-cell sample. For this reason, DARE may not perfectly de-

tect minor differences between cells at every single assayable CCGG site. However, it has been reported that functionally relevant methylation differences are generally associated with genomic regions rather than individual CpGs (42). By applying the tiling window approach commonly used in bisulfite-based single-cell methylome assays to scDARE (43), we show that coordinated DNA methylation differences between cells involving multiple CCGG sites are robustly detected at single-cell level using DARE assay (Figure 5G), pointing to the sensitivity of DARE for detecting differences between different cell populations.

In this work, we rely on manual single-cell isolation to perform scDARE in PCR tubes. However, the simple, single-tube DARE protocol is particularly amenable to high throughput single-cell isolation and processing workflows afforded by flow cytometry sorting and microfluidic solutions. Currently, scDARE enables multimodal measurement of genetic and epigenetic information in single cells, but it is expected to be compatible with reported techniques that separate mRNA from genomic DNA (25,44) to simultaneously profile and investigate the relationship between genetic, epigenetic and transcriptomic profiles in single cells.

LM-PCR based WGA has been shown to be compatible with fixed samples including formalin fixed paraffin embedded (FFPE) tissue samples (45). Future work will focus on optimizing DARE for processing fixed tissues, which will greatly expand its use for clinical samples. Finally, MSRE-based methylation profiling approaches have also found applications for investigating other DNA modifications such as 5-hydroxymethylation (5hmC) (46). We expect that the principles of DARE may be extended to simultaneously profile genome-wide unmodified cytosines, 5mCs and 5hmC in the same sample.

CONCLUSIONS

Recent discoveries of extensive genetic intratumor heterogeneity have sparked the development of various single-cell WGA and analysis technologies to better understand tumorigenesis and stratify patients for treatment. At the same time, emerging evidence point to epigenetic abnormalities as a hallmark of cancer, including cases of convergent genetic and epigenetic evolution in tumors and their metastatic subclones (47). Recent studies have also highlighted the functional relevance of DNA methylation profiles in single circulating tumor cells (CTCs) or CTC clusters that relates to metastatic potential (48). Hence, a technology that enables WGA to concurrently report on genome-wide DNA methylation states will bridge the gap and expand on the current genome-centric view of tumor heterogeneity

In this study, we established DARE as an adapter-linker PCR based WGA approach to detect copy number aberrations at 500kb resolution in low-input and single-cell samples, and provided proof-of-concept for detecting SNVs. At the same time, the combined use of methylation sensitive/insensitive restriction enzymes and methylation-specific adapters allow genome-wide determination of methylation state at ∼1.38 million CCGG sites. To validate the technique, we demonstrate the ability of DARE to distinguish cell-type specific CNA and DNA methylation profiles in single cells. We anticipate that DARE will enable a wide range of applications, including interrogating genetic and epigenetic heterogeneity in different spatial regions of tumor by combining tissue microdissection with low-input DARE, and multimodal profiling of isolated CTCs for precise molecular classification of cancer patients with scDARE.

DATA AVAILABILITY

Sequencing data have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE128560.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank William Burkholder and Stephen Quake for helpful initial discussions. We also thank Chadi EL Farran and Jonathan Loh for valuable discussion on bioinformatics analysis of DARE.

FUNDING

Young Investigator Grant [1510151027 to L.F.C.] from the Biomedical Research Council of the Agency for Science, Technology and Research in Singapore. Funding for open access charge: NUS Startup Grant.
Conflict of interest statement. None declared.

REFERENCES

1. Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med.*, **1**, 62.
2. Steeg,P.S., Ouatas,T., Halverson,D., Palmieri,D. and Salerno,M. (2003) Metastasis suppressor genes: basic biology and potential clinical use. *Clin. Breast Cancer*, **4**, 51–62.
3. Chimonidou,M., Strati,A., Tzitzira,A., Sotiropoulou,G., Malamos,N., Georgoulias,V. and Lianidou,E.S. (2011) DNA Methylation of tumor suppressor and metastasis suppressor genes in circulating tumor cells. *Clin. Chem.*, **57**, 1169–1177.
4. Deleye,L., Tilleman,L., Vander Plaetsen,A.-S., Cornelis,S., Deforce,D. and Van Nieuwerburgh,F. (2017) Performance of four modern whole genome amplification methods for copy number variant detection in single cells. *Sci. Rep.*, **7**, 3422.
5. Spits,C., Le Caignec,C., De Rycke,M., Van Haute,L., Van Steirteghem,A., Liebaers,I. and Sermon,K. (2006) Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.*, **1**, 1965–1970.
6. Arneson,N., Hughes,S., Houlston,R. and Done,S. (2008) Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). *CSH Protoc.*, **2008**, pdb.prot4919.
7. Zong,C. (2017) *Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) for the Analysis of DNA Copy Number Variation*. Humana Press, NY, pp. 133–142.
8. Binder,V., Bartenhagen,C., Okpanyi,V., Gombert,M., Moehlendick,B., Behrens,B., Klein,H.-U., Rieder,H., Ida Krell,P.F., Dugas,M. *et al.* (2014) A new workflow for whole-genome sequencing of single human cells. *Hum. Mutat.*, **35**, 1260–1270.
9. Ferrarini,A., Forcato,C., Buson,G., Tononi,P., del Monaco,V., Terracciano,M., Bolognesi,C., Fontana,F., Medoro,G., Neves,R. *et al.* (2018) A streamlined workflow for single-cells genome-wide copy-number profiling by low-pass sequencing of LM-PCR whole-genome amplification products. *PLoS One*, **13**, e0193689.

10. Farlik,M., Sheffield,N.C., Nuzzo,A., Datlinger,P., Schönegger,A., Klughammer,J. and Bock,C. (2015) Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.*, **10**, 1386–1397.
11. 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.
12. Guo,H., Zhu,P., Wu,X., Li,X., Wen,L. and Tang,F. (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.*, **23**, 2126–2135.
13. Suzuki,M., Liao,W., Wos,F., Johnston,A.D., DeGrazia,J., Ishii,J., Bloom,T., Zody,M.C., Germer,S. and Greally,J.M. (2018) Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Res.*, **28**, 1364–1371.
14. Ball,M.P., Li,J.B., Gao,Y., Lee,J.-H., LeProust,E.M., Park,I.-H., Xie,B., Daley,G.Q. and Church,G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.*, **27**, 361–368.
15. Oda,M., Glass,J.L., Thompson,R.F., Mo,Y., Olivier,E.N., Figueroa,M.E., Selzer,R.R., Richmond,T.A., Zhang,X., Dannenberg,L. *et al.* (2009) High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res.*, **37**, 3829–3839.
16. Suzuki,M., Jing,Q., Lia,D., Pascual,M., McLellan,A. and Greally,J.M. (2010) Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol.*, **11**, R36.
17. Han,L., Wu,H.-J., Zhu,H., Kim,K.-Y., Marjani,S.L., Riester,M., Euskirchen,G., Zi,X., Yang,J., Han,J. *et al.* (2017) Bisulfite-independent analysis of CpG island methylation enables genome-scale stratification of single cells. *Nucleic Acids Res.*, **45**, e77.
18. Andrews,S. (2010) *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
19. Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
20. Marcel,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJournal*, **17**, 10–12.
21. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
22. Kendall,J. and Krasnitz,A. (2014) Computational methods for DNA copy-number analysis of tumors. *Methods Mol. Biol.*, **1176**, 243–259.
23. Zhou,B., Ho,S.S., Greer,S.U., Spies,N., Bell,J.M., Zhang,X., Zhu,X., Arthur,J.G., Byeon,S., Pattni,R. *et al.* (2019) Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res.*, **47**, 3846–3861.
24. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
25. Hou,Y., Guo,H., Cao,C., Li,X., Hu,B., Zhu,P., Wu,X., Wen,L., Tang,F., Huang,Y. *et al.* (2016) Single-cell triple omics sequencing reveals genetic, epigenetic and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.*, **26**, 304–319.
26. Ning,L., Li,Z., Wang,G., Hu,W., Hou,Q., Tong,Y., Zhang,M., Chen,Y., Qin,L., Chen,X. *et al.* (2015) Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. *Sci. Rep.*, **5**, 11415.
27. Gravina,S., Dong,X., Yu,B. and Vijg,J. (2016) Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biol.*, **17**, 150.
28. Volkov,P., Bacos,K., Ofori,J.K., Esguerra,J.L.S., Eliasson,L., Rönn,T. and Ling,C. (2017) Whole-Genome bisulfite sequencing of human pancreatic islets reveals novel differentially methylated regions in type 2 diabetes pathogenesis. *Diabetes*, **66**, 1074–1085.
29. Barrera,V. and Peinado,M.A. (2012) Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale. *Nucleic Acids Res.*, **40**, 11490–11498.
30. Gu,H., Smith,Z.D., Bock,C., Boyle,P., Gnirke,A. and Meissner,A. (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.*, **6**, 468–481.
31. Mendizabal,I., Zeng,J., Keller,T.E. and Yi,S. V. (2017) Body-hypomethylated human genes harbor extensive intragenic transcriptional activity and are prone to cancer-associated dysregulation. *Nucleic Acids Res.*, **45**, 4390–4400.
32. Zilberman,D. (2017) An evolutionary case for functional gene body methylation in plants and animals. *Genome Biol.*, **18**, 87.
33. Jjingo,D., Conley,A.B., Yi,S. V, Lunyak,V and Jordan,I.K. (2012) On the presence and role of human gene-body DNA methylation. *Oncotarget*, **3**, 462–474.
34. Liu,Y., Siegmund,K.D., Laird,P.W. and Berman,B.P. (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.*, **13**, R61.
35. 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.
36. Zhou,B., Ho,S.S., Greer,S.U., Zhu,X., Bell,J.M., Arthur,J.G., Spies,N., Zhang,X., Byeon,S., Pattni,R. *et al.* (2019) Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.*, **29**, 472–484.
37. Li,T.-H., Kim,C., Rubin,C.M. and Schmid,C.W. (2000) K562 cells implicate increased chromatin accessibility in Alu transcriptional activation. *Nucleic Acids Res.*, **28**, 3031–3039.
38. Oda,M., Glass,J.L., Thompson,R.F., Mo,Y., Olivier,E.N., Figueroa,M.E., Selzer,R.R., Richmond,T.A., Zhang,X., Dannenberg,L. *et al.* (2009) High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res.*, **37**, 3829–3839.
39. Han,L., Wu,H.-J., Zhu,H., Kim,K.-Y., Marjani,S.L., Riester,M., Euskirchen,G., Zi,X., Yang,J., Han,J. *et al.* (2017) Bisulfite-independent analysis of CpG island methylation enables genome-scale stratification of single cells. *Nucleic Acids Res.*, **45**, e77.
40. Jelinek,J., Liang,S., Lu,Y., He,R., Ramagli,L.S., Shpall,E.J., Estecio,M.R.H. and Issa,J.-P.J. (2012) Conserved DNA methylation patterns in healthy blood cells and extensive changes in leukemia measured by a new quantitative technique. *Epigenetics*, **7**, 1368–1378.
41. Guo,H., Zhu,P., Wu,X., Li,X., Wen,L. and Tang,F. (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.*, **23**, 2126–2135.
42. Hansen,K.D., Langmead,B. and Irizarry,R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.
43. Angermueller,C., Clark,S.J., Lee,H.J., Macaulay,I.C., Teng,M.J., Hu,T.X., Krueger,F., Smallwood,S.A., Ponting,C.P., Voet,T. *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, **13**, 229–232.
44. Macaulay,I.C., Haerty,W., Kumar,P., Li,Y.I., Hu,T.X., Teng,M.J., Goolam,M., Saurat,N., Coupland,P., Shirley,L.M. *et al.* (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods*, **12**, 519–522.
45. Stoecklein,N.H., Erbersdobler,A., Schmidt-Kittler,O., Diebold,J., Schardt,J.A., Izbicki,J.R. and Klein,C.A. (2002) SCOMP is superior to degenerated oligonucleotide primed-polymerase chain reaction for global amplification of minute amounts of DNA from microdissected archival tissue samples. *Am. J. Pathol.*, **161**, 43–51.
46. Petterson,A., Chung,T.H., Tan,D., Sun,X. and Jia,X.-Y. (2014) RRHP: a tag-based approach for 5-hydroxymethylcytosine mapping at single-site resolution. *Genome Biol.*, **15**, 456.
47. Brocks,D., Assenov,Y., Minner,S., Bogatyrova,O., Simon,R., Koop,C., Oakes,C., Zucknick,M., Lipka,D.B., Weischenfeldt,J. *et al.* (2014) Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.*, **8**, 798–806.
48. Gkountela,S., Castro-Giner,F., Szczerba,B.M., Vetter,M., Landin,J., Scherrer,R., Krol,I., Scheidmann,M.C., Beisel,C., Stirnimann,C.U. *et al.* (2019) Circulating tumor cell clustering shapes DNA methylation to enable metastasis seeding. *Cell*, **176**, 98–112.