Low-input and single-cell methods for Infinium DNA methylation BeadChips

Sol Moe Lee [1], **Christian E. Loo**[2], **Rexxi D. Prasasya**[3], **Marisa S. Bartolomei**[3], **Rahul M. Kohli**[4] **and Wanding Zhou** [1,5,*]

[1]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, PA 19104, USA
[2]Graduate Group in Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Cell and Developmental Biology, Epigenetics Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
[4]Department of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed. Tel: +1 215 590 0473; Email: wanding.zhou@pennmedicine.upenn.edu

Abstract

The Infinium BeadChip is the most widely used DNA methylome assay technology for population-scale epigenome profiling. However, the standard workflow requires over 200 ng of input DNA, hindering its application to small cell-number samples, such as primordial germ cells. We developed experimental and analysis workflows to extend this technology to suboptimal input DNA conditions, including ultra-low input down to single cells. DNA preamplification significantly enhanced detection rates to over 50% in five-cell samples and ~25% in single cells. Enzymatic conversion also substantially improved data quality. Computationally, we developed a method to model the background signal's influence on the DNA methylation level readings. The modified detection *P*-value calculation achieved higher sensitivities for low-input datasets and was validated in over 100 000 public diverse methylome profiles. We employed the optimized workflow to query the demethylation dynamics in mouse primordial germ cells available at low cell numbers. Our data revealed nuanced chromatin states, sex disparities, and the role of DNA methylation in transposable element regulation during germ cell development. Collectively, we present comprehensive experimental and computational solutions to extend this widely used methylation assay technology to applications with limited DNA.

Graphical abstract
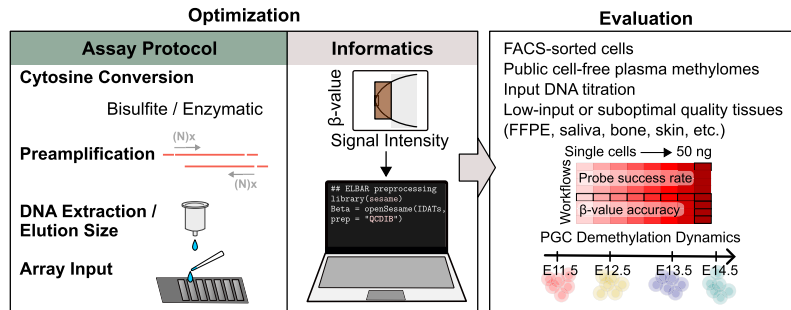
Introduction

DNA modifications, including 5-methylcytosines (5mCs) and 5-hydroxymethylcytosines (5hmCs), are canonical forms of epigenetic modification in human and other mammalian genomes. DNA methylation is found mainly in CpG dinucleotide contexts, where it is extensively implicated in gene transcriptional regulation, cell identity maintenance, organismal development, aging, and diseases (1). Infinium DNA methylation BeadChips are among the most popular genome-wide methylation assays in humans and other species due to the ease of experiment and data analysis (2). These arrays have been the primary data generation workhorse for large data consortia such as The Cancer Genome Atlas (TCGA),

with public methylome profiles of over 80 000 HM450 samples (3) and a similar number of EPIC array methylome profiles deposited to Gene Expression Omnibus (GEO). While the adoption of sequencing-based methods is catching up in case and mechanistic studies, the Infinium technology remains the most used assay platform for population-level studies such as meQTL studies (4,5), epigenetic risk scores (6,7), and other epigenome-wide association studies (8,9). This is, in part, due to the necessity in population studies to cover a large number of samples with nuanced variation in methylation levels and to dissect multiple cohort covariates such as sex, age, genetic background, and tissue type. In addition to being a powerful and popular tool for biological discovery, the technology

has recently enabled rapid clinical application development (10). This platform has found wide success in cancer diagnosis (11), cell-free liquid biopsy (12), and forensics (13). Recently, the Infinium BeadChips have also been used to generate the largest DNA methylome atlas across different mammalian species (14–16).

Despite these successes, a significant drawback of this technology is that it requires over 200 ng of input DNA from the standard processing protocol (2). This requirement constrains scientific and clinical applications with limited DNA availability. For example, as few as 25 primordial germ cells (PGCs) can be found in the early mouse embryo (17). Serum-derived, tumor-originated cell-free DNA (cfDNA) in cancer patients (18) holds value in non-invasive early cancer diagnosis (19). But as little as five ng/ml cfDNA in healthy subjects and 30 ng/ml in cancer patients (20) may be available. Also, DNA obtained from crime scene traces are often in the picogram to nanogram range (21). The DNA concentrations obtained from these tissues are much lower than the Infinium array standard protocol requires.

In addition to circumstances where the input DNA quantity is limited, it may be of interest to study a complex tissue to dissect cell-to-cell heterogeneity purposefully, even when there is no shortage of total DNA in the tissue. DNA methylation encodes distinctive cell identity fingerprints, which can be used to infer cellular phenotypes (1), trace the state of the DNA-releasing cells (22), and infer cell proportions (23). By performing DNA methylation analysis on laser-capture microdissected specimens, one can compare methylomes at different locations in tumors (24,25) or select specific cell types from the brain for analysis (26). In most cases, laser-capture microdissected tissues are limited in quantity and involve pre-assay whole genome amplification, as is done with SNP arrays (27).

The extreme of increasing the cell resolution at the cost of working with small DNA amounts is epitomized by the rapid development of single-cell DNA methylome assay technologies (28). Each human or mouse cell carries 5–7 pg. In the past decade, most technologies have been based on pre-amplifying deaminated DNA (29–31) using random priming or single-strand adapter ligation before feeding amplified DNA to high-throughput sequencing. In addition, multiple enzymatic cytosine conversion methods have been developed to replace sodium bisulfite conversion to better preserve genomic DNA during library preparation (32). Inspired by these single-cell methods, we posit that similar preamplification methods and enzymatic conversion can also be used with Infinium arrays.

Besides the experimental challenges, current computational methods are not fully optimized for low-input DNA data. Current signal preprocessing practices may lead to low probe signal detection rates and probe over-masking. Signal detection has traditionally been determined by comparing probes' signal intensities with negative control probes (33) or readings in the Infinium-I out-of-band channel (34). A conservative threshold of the detection *P*-value was then determined to mask low-intensity probes. When the foreground and background signal intensities are separable, using a conservative threshold on high-input data would not harm detection sensitivity. This approach leads to a significant loss of true biological signals for limited DNA input. A method that maximally preserves biological signals while removing pure artifacts remains an unmet need.

Here, we systematically developed and evaluated experimental and computational methods to improve array sensitivity at low-input ranges and single cells. Our evaluation encompasses previously attempted adaptations, including using different bisulfite conversion elution (35), using Formalin-Fixed Paraffin-Embedded (FFPE) restoration (36), combining bisulfite conversion with DNA extraction (37) as well as methods that were never previously used with Infinium arrays, such as the enzymatic conversion and different preamplification strategies. We developed a new signal detection framework to address the computational challenge of processing data from limited DNA. We showed that this new method significantly improved array detection rates while effectively masking probes whose readings are dominated by background signals. We showed that the Infinium BeadChip is compatible with samples of low input down to single cells. And we presented end-to-end solutions to enable this technology for low-input and single-cell samples.

Materials and methods

Cell cultures

NIH3T3 (ATCC, CRL-1658) was obtained from American Type Culture Collection (ATCC) and cultured in DMEM (ATCC, 30-2002) containing 10% Calf Bovine Serum (ATCC, 30-2030) and 1% penicillin/streptomycin (Gibco, 15140122). B16-F0 (ATCC, CRL-6322) was obtained from ATCC and cultured in DMEM (ATCC, 30-2002) containing 10% Fetal Bovine Serum (Gibco, 45000-736) and 1% penicillin/streptomycin (Gibco, 15140122). All cells were maintained in a 37°C incubator with 5% CO_2 and cultured in a 75 cm^2 culture flask (Fisher, BD353136).

Cell flow sorting

5×10^6 cell pellets of the NIH3T3 and the B16-F0 were resuspended in 50 μl of 0.1 μg/1 ml of 4,6-diamidino-2-phenylindole (DAPI) (Sigma-Aldrich, D9542-5MG) in 1 ml of phosphate-buffered saline (PBS) (Life Technology, 10010023). Cells were filtered by a Falcon Cell Strainer Snap Cap (Falcon, 352235). DAPI-negative cells (1, 2, 5, 10 and 100 cells) from NIH3T3 and B16-F0 were sorted and collected into 96-well plates pre-loaded with 10 μl of 1× M-Digestion Buffer (Zymo Research, D5020-9) using a BD FACSAria™ Fusion cell sorter (BD Biosciences) with a 100 μm nozzle.

Mouse primordial germ cells

Gonads from embryonic Oct4-GFP transgenic mice (B6;129S4-*Pou5f1^{tm2Jae}*/J; Jackson Laboratory, strain #008214,RRID: IMSR_JAX:008214) were harvested at embryonic day E11.5, E12.5, E13.5, and E14.5 (38). Gonads were dissected in calcium- and magnesium-free PBS (Gibco) and transferred into 500 μl of 0.25% Trypsin–EDTA (Gibco). Subsequently, the preparation of embryonic germ cells was carried out following the method previously described (39). For bisulfite mutagenesis, PGCs were snap-frozen for storage at –80°C until further processing.

DNA extraction and bisulfite conversion

NIH3T3 and B16-F0 cells were harvested by centrifugation at 100g for 5 min at room temperature and washed twice using PBS (Gibco, 10010023). The DNeasy Blood and

Tissue Kit (Qiagen, 69504) was used to extract genomic DNA from NIH3T3 and B16, according to the manufacturer's protocol. DNA samples were quantified using Qubit 4.0 Fluorometer (Invitrogen) using the dsDNA HS Assay Kit (Invitrogen, Q33231). Bisulfite conversion was performed using three kits. DNA bisulfite conversion using EZ DNA Methylation Kit (Zymo Research, D5001) was performed according to the manufacturer's instructions with the specified modifications for Illumina Infinium Methylation Assay. DNA bisulfite conversion using EZ DNA Methylation-Gold Kit (Zymo Research, D5005) and EZ DNA Methylation-Direct Kit (Zymo Research, D5020) was performed according to the manufacturer's protocol. Cell lysis and bisulfite conversion from sorted cells and PGCs were performed with EZ DNA Methylation-Direct Kit according to the manufacturer's instructions.

DNA restoration

After bisulfite conversion, the bisulfite-converted DNA was eluted, resulting in an 8 µl volume. The Infinium HD FFPE DNA restoration kit (Illumina, WG-321-1002) was then used according to the manufacturer's instructions. Following a 1-min incubation, the elution of the DNA was carried out using autoclaved ultrapure water for a 10 µl elution volume. The eluted DNA was stored at –20°C before undergoing Infinium array processing. Intermediate DNA purifications were performed using the DNA Clean and Concentrator-25 Kit (Zymo Research, D4064).

Cytosine conversion elution size optimization

The Illumina Infinium Mouse Methylation BeadChip assays were conducted according to the manufacturer's specifications with slight modifications (Supplementary Tables S1A and S1B). The original protocol specifies using 4 µl obtained from 12 to 22 µl eluted BCD, mixed with 4 µl 0.1 N sodium hydroxide for amplification and BeadChip reaction (Infinium HD Assay Methylation Protocol Guide 15019519 v01). However, commercial bisulfite conversion kits typically produce over 10 µl elution in purifying the converted DNA, leading to only part of the eluted DNA (4 µl) used for the BeadChip assay. Previous studies have adjusted the elution size or additional concentration steps to minimize DNA loss. For example, one option is to mix 7 µl eluted DNA with 1 µl 0.4 N sodium hydroxide (NaOH) (40–42). DNA input can be maximized by increasing NaOH concentration in denature step of the Infinium array. To preserve more input DNA, we compared four alternative combinations of elution buffer, input DNA volume, and NaOH concentration and volume.

Enzymatic methyl (EM)-array sample preparation

Libraries were prepared using the NEBNext Enzymatic Methyl-seq (NEB, E7120S) kit, following the manufacturer's instructions. 50, 5, 2, or 0.5 ng of 5mC adaptor-ligated NIH3T3 DNA was used as input. HiFi HotStart Uracil + Ready Mix (KAPA Biosystems, KK2801) was used to amplify the libraries following conversion before purification over SPRI beads (0.8× left-sided) and elution in nuclease-free water to yield final libraries. Libraries were then quantified by Qubit HS (Invitrogen, Q32851) and quality checked on an Agilent Bioanalyzer 2100 before sequencing on an Illumina MiSeq instrument to confirm conversion efficiencies. Details of each input's elution size and library amplification cycles are listed in Supplementary Table S1C.

ELBAR detection *P*-value calculation

In the standard Infinium BeadChip usage, >200 ng DNA is profiled, and probe detection calling is employed to filter out probes whose signals are subject to substantial background influence. However, this practice will cause significant biological signal loss from low-input datasets where users often seek to retain the most biological signal but can tolerate some background influence. To meet this need, we developed ELBAR (Eliminating BAckground-dominated Reading) to exclude/mask only probes lacking biological signals and entirely dominated by background signals. The ELBAR method is based on the observation that the beta value ranges depend on the probes' total signal intensities and includes the following steps. First, we define total signal intensities as the sum of signals methylated (M) and unmethylated (U) alleles. Pooling in-band and out-of-band signals, ELBAR bins probes by $\log_2$-transformed M + U signal intensities. Next, ELBAR calculates the upper and lower bounds (calculated using the 5% and 95% quantiles to accommodate outliers, defined as the *beta value envelope*) of each bin as M + U varies. Third, we define the background signal by looking for the first bin that deviates in the beta value envelope of the bin from the smallest M + U. Lastly, these probes' maximum M and U signals were treated as the true background signal to compute detection *P*-values.

Public datasets

The mouse MM285 datasets were downloaded from GEO under GSE18441 (15,43). GEO accessions of other EPIC and HM450 datasets were provided in Supplementary Table S2A. BS-seq dataset for 4 PGC samples (E10.5, E11.5, one male E12.5, and one female E12.5) was downloaded from GEO under GSE76971 (44). TCGA testicular seminoma datasets (45) were downloaded from Genomic Data Commons (46).

Infinium BeadChip data preprocessing and analysis

The Illumina Infinium BeadChips technology is based on sodium bisulfite conversion of DNA, with single base resolution genotyping of targeted CpG sites via probes on a microarray. Probes are designed to match specific 50 base regions of bisulfite-converted genomic DNA, with a CpG site at the probe's 3′ end (47). Upon hybridizing with bisulfite-converted DNA, the probe undergoes a single-base extension that incorporates a fluorescently labeled ddNTP at the 3′ CpG site, enabling distinguishing of the C/T conversion resulting from bisulfite conversion. The fluorescent signals provide insight into the methylation status (methylated or unmethylated) of specific cytosine residues in the DNA sample. Signal intensity refers to the strength of the fluorescent signal emitted by the hybridized probes on the BeadChip. The signal intensity is directly related to the quantity of target molecules bound to specific probes. Probe success rates represent the proportion of successfully captured CpG probes for targeted CpGs.

The IDAT files generated, along with all public datasets utilized, were processed using the SeSAMe R package. This encompassed preprocessing, quality control, and analysis, adhering to the established preprocessing workflow (43). The probe detection *p*-value was computed using the pOOBAH algorithm, which leverages the fluorescence of out-of-band (OOB) probes. Subsequently, normalization was performed using noob, which applies a normal exponential deconvolution of fluorescent intensities based on the OOB probes.

Additionally, a dye bias correction was applied using the dyeBiasNL function. Infinium Methylation BeadChip manifest and annotations data, which include gene, chromatin state, sequence context, the 'PGCMeth' probes list (designed to target CpGs highly methylated in E13.5 PGCs) (48), and other functional annotations, were obtained from http://zwdzwd.github.io/InfiniumAnnotation. Metagene plot was generated using the KYCG_plotMeta function in the SeSAMe package. To calculate the F1 scores of each sample against the 250-ng control (205243950081_R01C01), a beta value greater than 0.5 was rounded to one and set to zero otherwise. Then, the F1 score is calculated by treating one as true and 0 as false and comparing the target sample with the 250-ng control, i.e. $F1 = 2TP / (2TP + FN + FP)$ where TP is the true positive counts, FP is the false positive counts, and FN is the false negative counts.

Results

Characterizing the suboptimal DNA signatures in public Infinium datasets

Suboptimal DNA quality and quantity impact Infinium methylation data through the manifestation of lower signal intensity and probe success rate. We first studied the probe success rate of over 100 000 public Infinium datasets deposited to GEO to identify the determinants of Infinium array performance (Figure 1A). Comparing DNA sources, we observed that bone, buccal, plasma cfDNA, esophagus, and saliva often yield data with suboptimal probe detection success. Further dataset stratification by sample preservation reveals that FFPE samples are significantly lower in probe success rate than non-FFPE samples. The lower signal intensity leads to skewing of beta values, which represent the methylation level at a specific site, towards an intermediate reading at 0.5 due to stronger relative influences of the signal background (Figure 1B–D) (34). This asymptotic convergence to 0.5, defined by the upper and lower bounds of beta values, forms a *beta value envelope*. This increase in the background signal influence is a continuous spectrum instead of a dichotomy of detection success vs. failure. When the input DNA is of high quality and quantity, most probes have stable and clustered signal intensities, leading to a bimodal beta value distribution. But in suboptimal datasets such as from FFPE and cfDNA samples, more probes carry lower signal intensities, leading to beta values approaching 0.5 (Figure 1C). This drop in sample quality in FFPE samples is likely intertwined with low DNA input amount, as supported by a similar transition of intermediate methylation readings in the low input data (Figure 1D and Supplementary Figure S1A). We also found that FFPE and cfDNA samples lose detection at different genomic regions. cfDNA, saliva, and buccal cells preferentially lose detection at GC-rich promoter sites such as TssA and TssBiv, while FFPE samples are less biased across genomic territories (Figure 1E). The full names of the chromatin states are available in Supplementary Figure S1B.

Infinium BeadChip workflows for low DNA input and single cells

To improve signal detection in low-input experiments, we developed 13 non-standard workflows using: (i) preamplification of the genomic DNA (see Supplementary Methods); and (ii) enzymatic cytosine conversion methods, besides other protocol adjustments that preserve DNA load (see Figure 2A,

Supplementary Tables S1A–S1C). The probe success rate and F1 scores were presented in Figure 2B–D for three workflows, each representing distinct characteristics: the unmodified workflow (Workflow A), the workflow with maximized elution size (Workflow C), and the most optimized workflow with preamplification (Workflow J). Additionally, Figure 2C and D included a workflow achieved through enzymatic base conversion, labeled as Workflow M. In Supplementary Figures 2B–2D, we illustrated probe success rates, F1 scores and Spearman correlations for all 13 workflows, encompassing the four previously mentioned workflows. We found that the standard Illumina workflow (Workflow A) can detect signals on 70% array probes with two ng DNA without modification, consistent with our prior characterization of the EPICv2 BeadChip (49). In the sub-2-ng range, the detection success rate drops rapidly for Workflows A and B (Figure 2B, C, and Supplementary Figure S2B). Enzymatic base conversion (Workflow M) maximizes signal detection (84%) in 0.5 ng-input experiments, followed by a whole-genome preamplification-based method (Workflow J) and one with elution size modification alone (Workflow C). Due to the allelic nature of DNA, we use the F1 score (Materials and methods), which binarizes beta values for comparison to the reference sample, rather than correlation. In the five-cell experiments, Workflow J (Figure 2B and D, and Supplementary Figure S2C) consistently reaches over 70% in signal detection (pOOBAH < 0.05) and close to 90% in the F1 score. Notably, Workflow J detects around 25% probes in single cells with an F1 score >70%, suggesting that the detected data is biologically informative despite a higher rate of detection failure. In contrast, the standard workflow is consistently under 50% in the F1 score, suggesting more biologically misleading readings.

For 0.5 and 2 ng input ranges, Workflow J also improved the data quality as indicated by the higher correlation coefficient with a 250-ng dataset (Figure 2E and F, and Supplementary Figure S2D). The improvement is most evident in the recovery of intermediate DNA methylation, which is only observed in workflow J at 0.5 ng (Figure 2F), suggesting that preamplification helps all samples from fewer than ten cells to those with over 1 ng input (Figure 2E).

Previous studies have shown that FFPE restoration kits may improve signal detection on DNA of suboptimal quality (36). The FFPE restoration kit did not significantly improve the array performance with 50 ng and 0.5 ng non-FFPE DNA input (Workflows D and E, Supplementary Figure S2E). We also did not observe substantial performance differences among the three different bisulfite conversion kits (Supplementary Figure S2F). EZ-direct kit produced data of a slightly better probe detection rate likely due to the minimization of DNA loss from a single-step bisulfite conversion and purification.

We also compared preamplification based on random hexamers (N6) and random hexamers with a T7 primer at the 5'-end (N6-T7) for whole genome amplification (Supplementary Figures S2F and S2G) (50). Workflow J showed the highest Spearman`s rho compared to other workflows with preamplification (Workflows F–I). The T7 sequence in N6-T7 serves as a second primer, allowing further PCR amplification. However, additional PCR cycles did not improve array performance (Workflows K and L, Supplementary Table S1D, Supplementary Figures S2B, and S2D). Moreover, the array with four Klenow amplification cycles (Workflow H) did not outperform the array with two cycles. Given our findings,
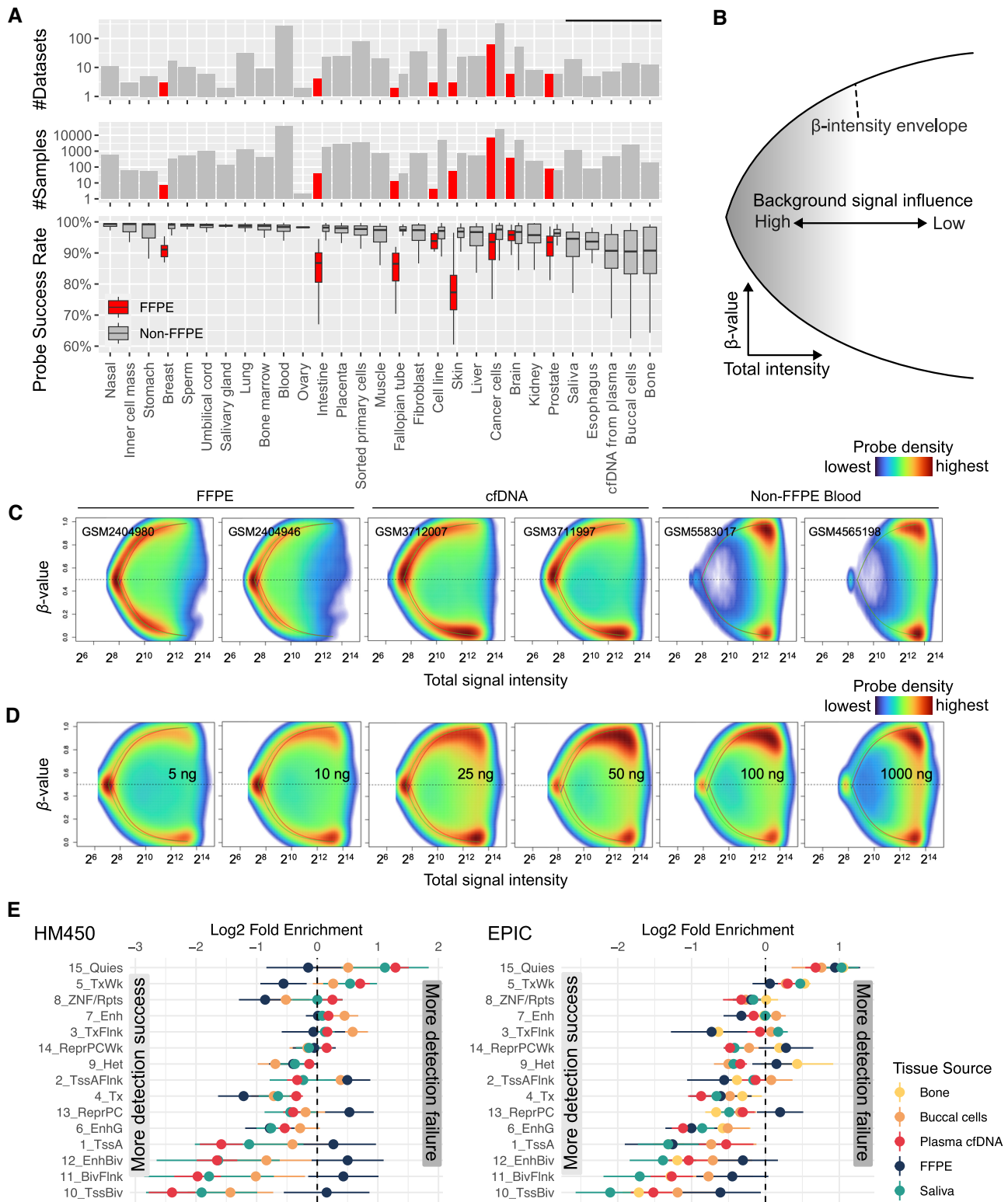
Figure 1. Lower input arrays exhibit lower signal intensity and lower probe success rates. (**A**) The probe success rate in 105475 public Infinium methylation BeadChip array data sets. (**B**) Schematic showing the dependence of beta value on the probe's total signal intensity. The x-axis represents the total signal intensity, and the y-axis represents beta values. (**C**) The intensity-beta plot compares different DNA sources and (**D**) input amounts in ng. The green and red curves denote the beta value envelopes defined by the green and the red channel background signal, respectively. (**E**) The probe detection rate of FFPE and cfDNA samples at different genomic regions from HM450 (left) array data and (right) EPIC array data.
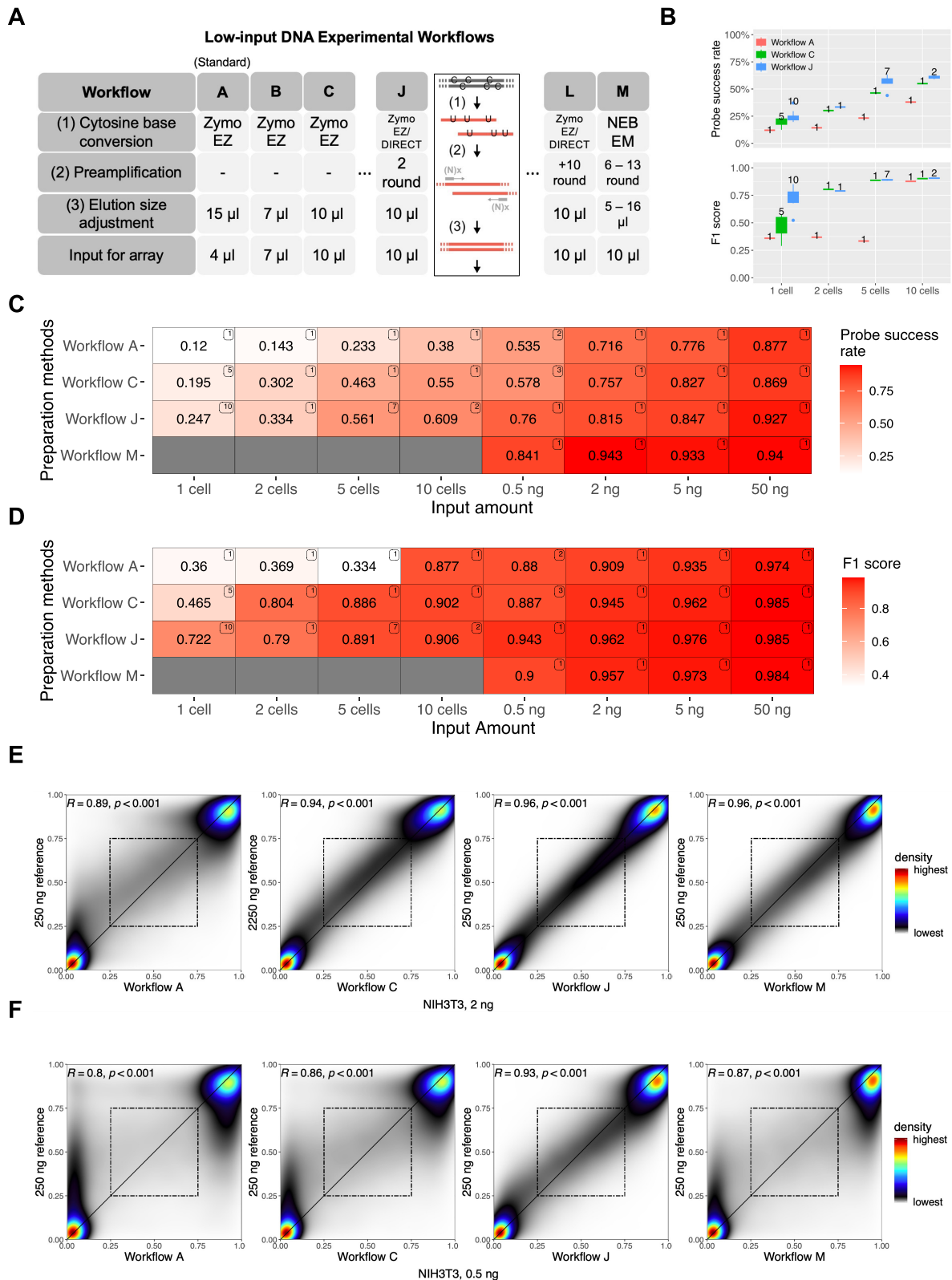
Figure 2. Infinium BeadChip performance in ultra-low input ranges. (**A**) A summary table of workflows used in this study. Workflow A is the Illumina standard workflow. (**B**) Box plots were used to visualize the probe success rates (top) and the F1 score (bottom). The number of samples for each experiment was displayed next to each box. (C, D) Comparison of four main preparation methods based on (**C**) probe success rate and (**D**) F1 score. The number on the top right corner of each tile indicates the number of samples analyzed in each experiment. See also Supplementary Figure S2. (E, F) Smooth scatterplots for the comparison of workflows A, C, J, and M with (**E**) 2 ng and (**F**) 0.5 ng of DNA input (*R*: spearman's rho, *P*: *P*-value). The dashed squares indicate intermediate beta values (0.25–0.75) on both axes.

we followed the N6 amplification strategy for the subsequent analysis.

Optimized workflow resolves intercellular heterogeneity while maintaining cell line identity

DNA methylomes profiled from a small number of cells often reveal cell-to-cell heterogeneity. We next tested whether our low-input method can uncover this heterogeneity and whether the cell population averages reduce to measurements from high DNA input. We merged the single and five-cell methylomes, respectively, and compared the combined data with the 250-ng methylome (Figure 3A). We found that the merged methylome reinstated the intermediate methylation measurements which are otherwise missing from the single low-input experiment representatives. As DNA methylation readouts are allelic (taking only 0%, 50%, and 100% in diploid cells), we expect a reduction of non-allelic fractions as the cell population becomes less heterogeneous. Focusing on CpGs showing intermediate methylation (0.3–0.7) in the 250-ng data, we observed a gradual dichotomization of methylation levels approaching 0% and 100% in the 10-cell ($n = 4$) datasets and more in the five-cell ($n = 5$) datasets (Figure 3B). This polarization is likely due to a higher genomic DNA homogeneity from reduced cell numbers. The lingering non-allelic methylation fractions are likely due to amplification bias and residual cell-to-cell heterogeneity. Given the same input cell number, preamplification (Workflow J) retained more intermediate readings (Figure 3C). The standard workflow (Workflow A) became nearly fully dichotomized at ten cells and struggled globally in the 5-cell experiment (Figure 3C).

Despite the resolution of intermediate methylation levels, small-cell-number data largely retain global and focal methylation differences intrinsic to the cell type identity (Supplementary Figure S3A). All datasets from five to ten cells cluster with the 250 ng datasets of the same cell line on a t-distributed stochastic neighbor embedding (tSNE) projection (Figure 3D). Most single-cell samples are also grouped accordingly despite the erroneous placement of two single-cell B16 samples, likely due to lack of biological signal, as suggested by their ultra-low probe detection rates (0.13) compared to other samples of similar input DNA amounts. The five-cell datasets are clearly separated in a metagene plot that suggested a higher global methylation for the B16 cells at all input ranges (Supplementary Figure S3B). Finally, the differentially methylated CpGs between the two cell lines are largely preserved in five-cell methylomes (Figure 3E). Random discrepant methylation did occur more frequently at CpGs intermediately methylated in the 250 ng datasets for the two cell lines, respectively (Figure 3E), enriching for bivalent chromatin (Supplementary Figures S3C and S3D). Collectively, these data suggest the Infinium arrays can robustly profile five-cell samples. While the Infinium arrays can profile single cells, their performances are unstable (Figure 3D).

GC-rich and high copy number regions retain detection in low input datasets

We next explored which genomic regions are most susceptible to detection loss in high and low input experiments. We first observed that signal intensities intrinsically depend on the probe sequences. We studied the within-sample intensity Z-score of HM450 autosomal probes across 749 normal samples from the TCGA cohort (Methods) (Figure 4A,

Supplementary Figure S4A). The violin plot displays the complete set of Z-scores for probes in the cohort. Data points corresponding to individual samples were smoothed and provided a comprehensive view of the overall distribution pattern. The Z-score distribution for different probes shows clear probe dependence. Probes at the high and low signal intensity extremes have little overlap, suggesting a strong sequence dependence. Probes targeting GC-rich regions (as indicated by the number of 'C's in the probe sequences since 'G's are replaced by 'A's to pair with 'T's from bisulfite conversion) are associated with higher signal intensities (Figure 4B). This is supported by an enrichment of the detected probes in CpG islands (Figure 4C, D, and Supplementary Figure S4B), gene promoters, transcription factor binding sites, e.g. TFAP2C, and promoter-associated histone modifications, e.g. H3K64ac (Supplementary Figures S4C and D).

Consistently, the probes that fail detection in 250-ng and 5-cell samples are significantly enriched in low-CpG density regions (Figure 4C). PMD solo-WCGWs—CpGs flanked by A/Ts and with no other CpGs within a 70-bp neighborhood at partially methylated domains (51)—are observed to lose most signal detection, consistent with their CpG-sparse nature. Probes targeting non-CG cytosine methylation also tend to lose detection in low-cell-number samples. Interestingly, mitochondrial CpG probes, transposable element (TE) CpGs, and other multi-mapping probes have the least probe detection loss in low-input datasets. The mitochondrial genome showed nearly 100% probe detection success in single, and two-cell experiments. This is likely due to the high copy number of mitochondrial genomes per cell (52). Similarly, other high copy number repetitive elements, such as the Satellite, B1 elements, and other SINE1/7SL elements, also show high probe success rates (Figure 4E and F). These results suggest that the multi-mapping probes may be used as a TE profiling tool for low-input samples. More prevalent heterogeneity in DNA methylation within quiescent chromatin was observed. CpGs displaying lower methylation levels in bulk tissues but higher levels in individual five cells (Figure 4G, right panel), or vice versa (Supplementary Figure S4E, right panel), are both enriched in quiescent chromatin regions. In contrast, promoters (Tss) and gene bodies (Tx) showed consistently low (Figure 4G, left panel) and high methylation levels (Supplementary Figure S4E, left panel), respectively, in both the bulk tissue and the 5-cell samples.

ELBAR preserves more signal detection for low-input datasets

The conventional detection *P*-value calculation aims at preventing false discovery in high-input datasets, where probes with suboptimal signal intensity are rare and a relatively clear decision boundary can be found. In low-input samples, more probes carry lower signal intensity and can overlap with measurements purely dominated by background signals (Supplementary Figure S5A). Applying the same detection *P*-value threshold may lead to a significant loss of biologically useful readings. To better balance sensitivity for low-input datasets, we developed the ELBAR algorithm, based on the observation that probes dominated by signal-background-only are always associated with intermediate methylation readings (Figure 5A). In brief, ELBAR looks for low-signal probes with intermediate methylation to model the background signal (Methods). Doing so can effectively remove
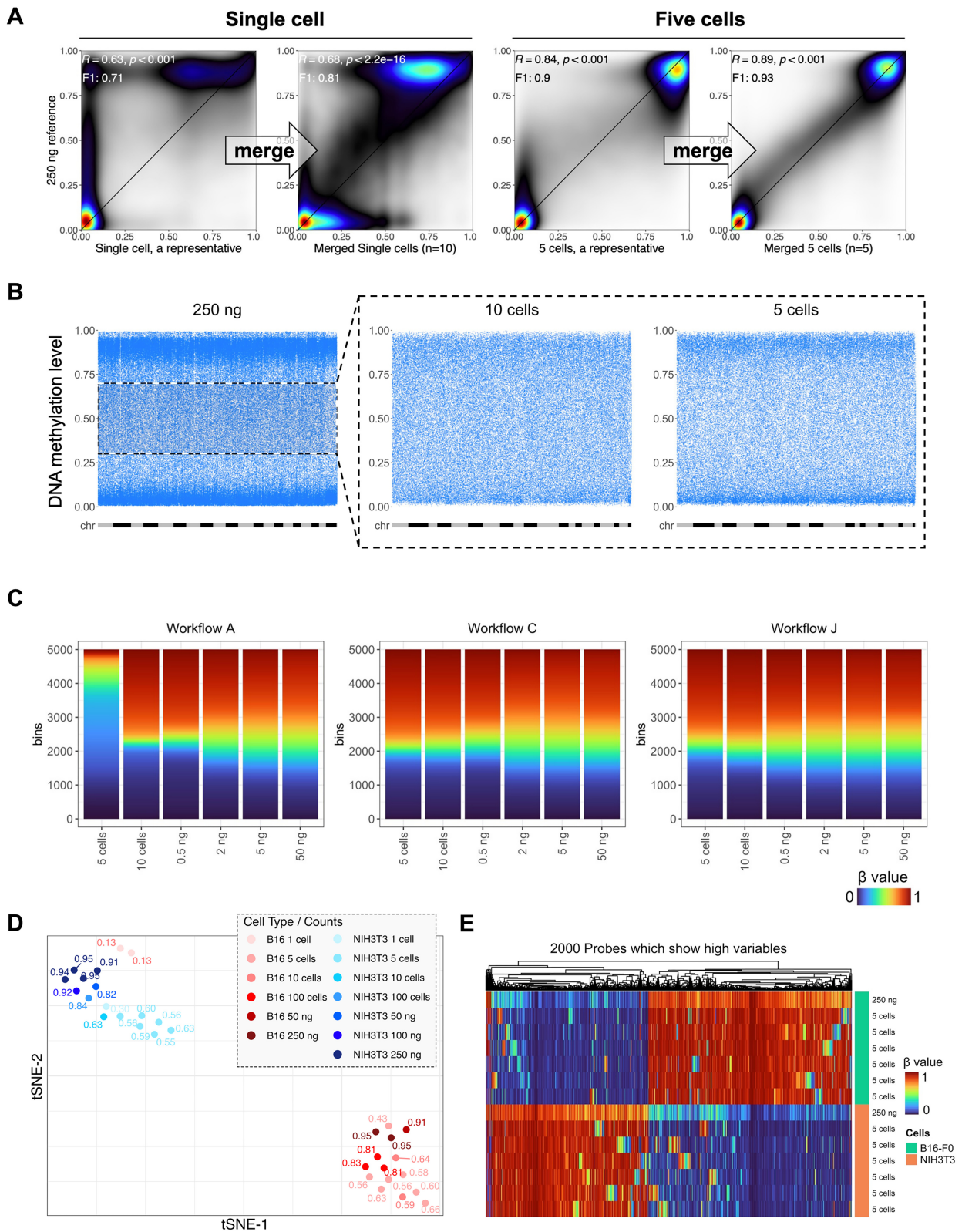
Figure 3. Low input methylation array data uncovered sample-to-sample heterogeneity. (**A**) Smooth scatter plots for methylomes from a representative single-cell dataset, the merged single-cell dataset ($N = 10$), a representative 5-cell dataset, and the merged five-cell dataset ($N = 5$), respectively, against the 250-ng input. (**B**) Distribution of intermediate methylation (0.3–0.7, dashed black square) of CpGs in the 250-ng methylome in 10- and 5-cell datasets. Each blue dot represents a CpG, and the X-axes were arranged in the order from chromosome 1 to chromosome Y, with each chromosome displayed in alternating gray and black colors. (**C**) DNA methylation level distribution of 5 cells to 50 ng with different workflows. (**D**) tSNE cluster of NIH3T3 and B16 using 2000 probes most variable in the methylation level. Fractions labeled in the plot are probe success rates for each dataset. All B16-F0, NIH3T3, and PGCs except 50, 100 and 250 ng were whole genomes amplified by Workflow I or J. (**E**) Heatmap showing array performance comparison of 5-cell methylomes of B16-F0 and NIH3T3 with 250-ng methylomes. 2000 highly variable CpGs with the highest standard deviations of DNA methylation value among samples were used.
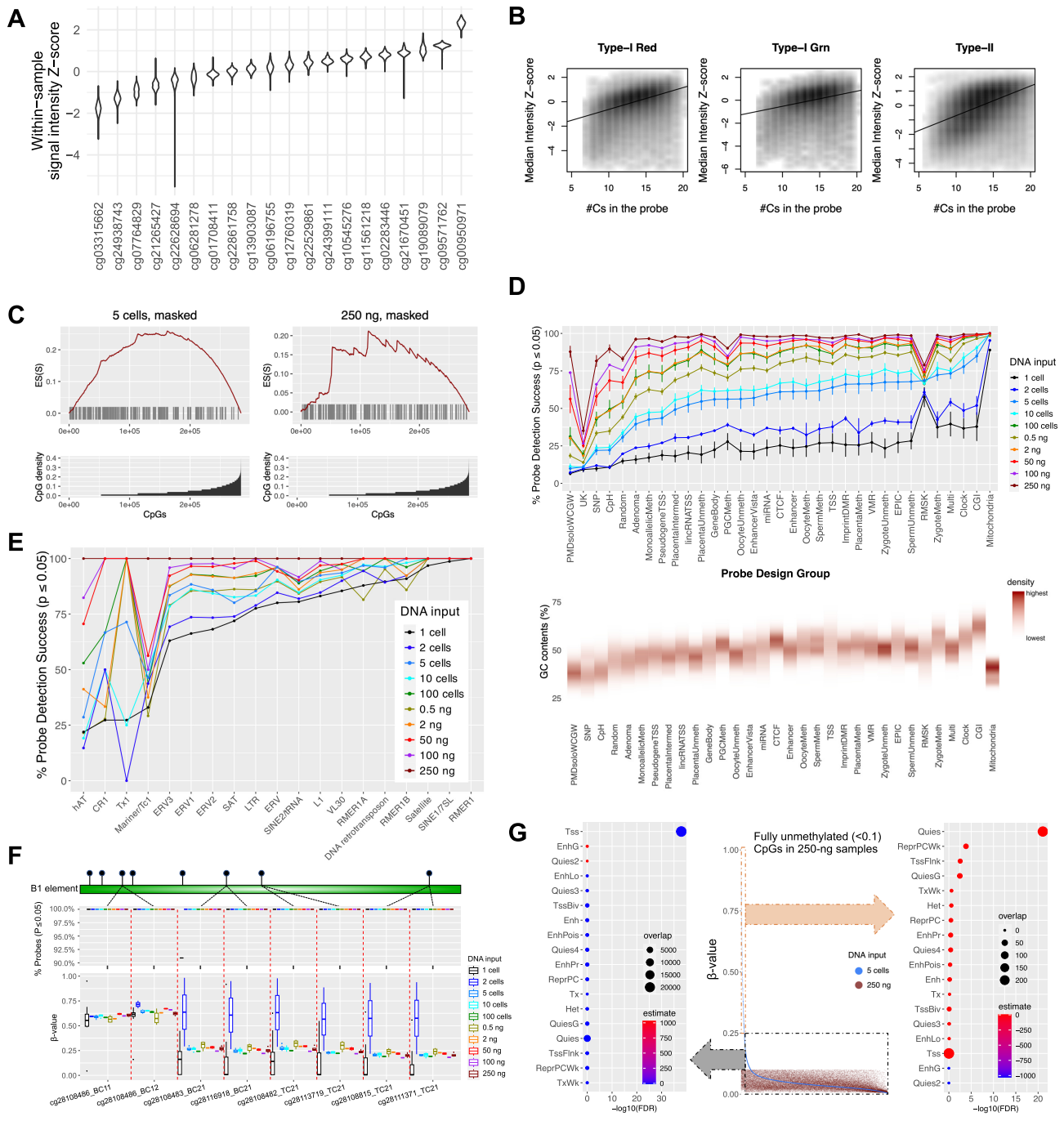
Figure 4. Low-input Infinium array preferentially loses detection on GC-sparse genome but retains detection on mitochondria and TEs. (**A**) Violin plot for intensity *Z*-score of representative HM450 autosomal probes across 749 normal samples from the TCGA cohort. (**B**) Correlation of signal intensities with the number of Cs in the probe sequence. (**C**) CpG density enrichment analysis for NIH3T3 5-cell and 250-ng datasets. (Bottom) The x-axis represents probes ($N = 297\,415$) organized in ascending order based on their CpG density. Y-axis indicates the CpG density for each probe. (Top) The x-axis aligns with the bottom plot, and the y-axis exhibits the enrichment score (ES). Each bar in the figure depicts the location of a pOOBAH-masked probe, arranged in ascending order of CpG density within a ±75 bp range of the flanking region surrounding the CpG target for each probe. (**D**) The mean and standard deviation (with error bars) of the probe success rate of probe design groups (a complete description of the definition of each probe design group is listed in the legend of supplementary Figure S4) (top). Flanking GC content ratios of probes on mouse methylation array within a range of ± 50 bp, categorized by probe design groups (bottom). (**E**) the mean of the probe success rate of TEs for varying DNA input amounts, ranging from single-cell to 250 ng datasets. The x-axes are organized in ascending sequence according to the probe success rates for five cells in (D) and single cells in (E). (**F**) (top) The probe detection rates of eight CpG probes targeting B1 elements ranged from single cells to 250 ng, and (bottom) corresponding beta values cover the same range. (**G**) The beta value distribution of CpGs in the 5 cells of NIH3T3 with blue dots, emphasizing fully unmethylated (less than 0.1 of beta values) in the 250-ng sample, represented by red dots (center). The x-axis was organized in descending order of beta values of the 5 cells. Dot plots for enriched chromatin states (75) for CpGs within the dashed gray square with delta beta less than 0.25 between the two samples (left), and CpGs within the dashed orange square with delta beta greater than 0.25 (right). Estimate: $\log_2$ of Odds ratio; overlap: number of curated CpG probes for each term of chromatin state.
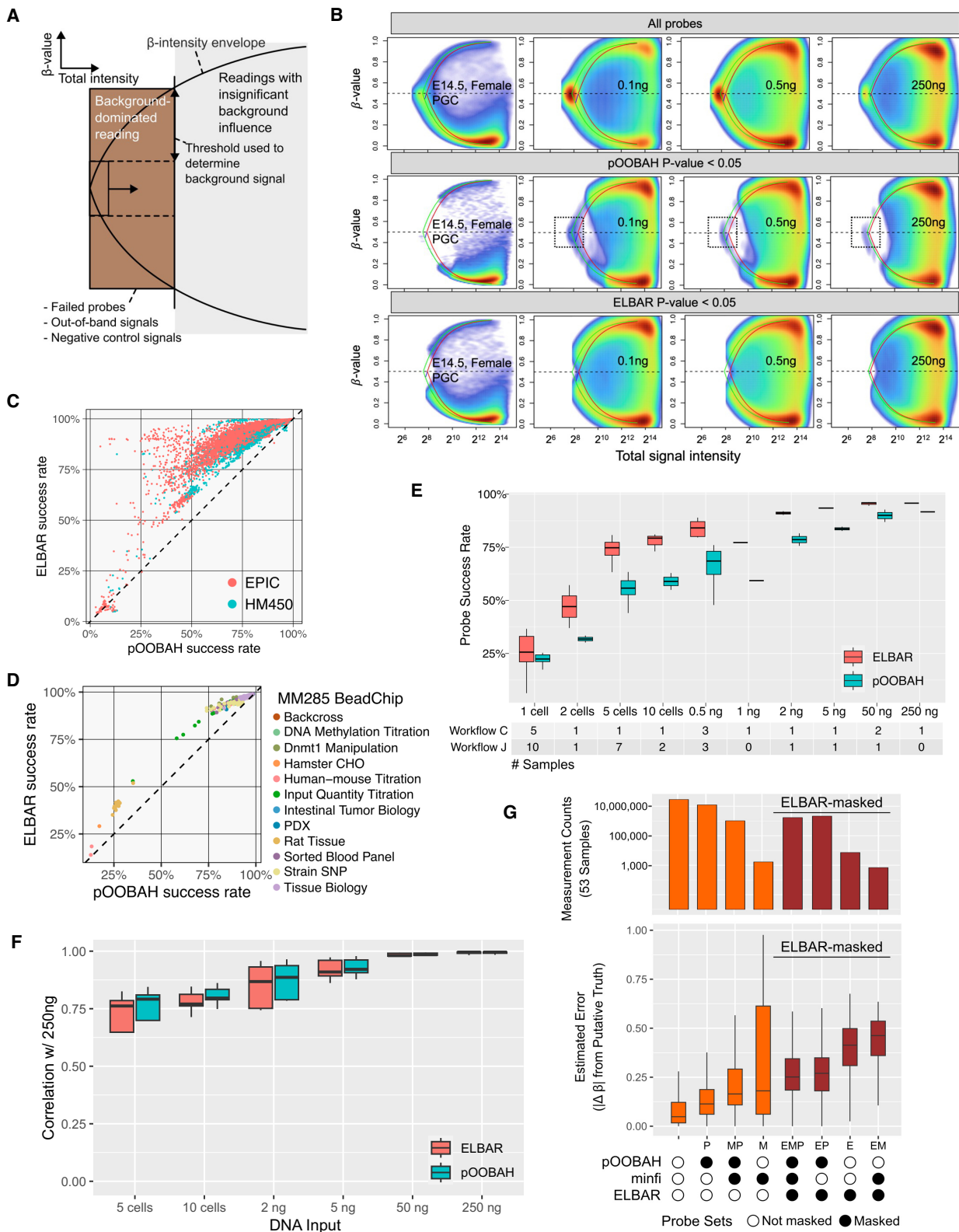
Figure 5. ELBAR detection preserves more probes with biological signals. (**A**) A schematic illustration of the ELBAR algorithm. The x-axis represents the total signal intensity, and the y-axis represents beta values. (**B**) Performance of ELBAR in eliminating background-dominated reading probes compared to pOOBAH and unfiltered in PGCs, 0.1-, 0.5- and 250-ng-input datasets. Dashed boxes illustrate the artificial, background-dominated readings left by pOOBAH masking. The green and red curves denote the beta value envelopes defined by the green and the red channel background signal, respectively. (C, D) ELBAR performance regarding the probe success rates for public human (**C**) and mouse (**D**) array datasets. (E, F) Comparison of ELBAR performance with pOOBAH regarding probe success rate (**E**, *P*-value = 2.7E-8, *t*-test of method difference in a multiple linear regression) and Spearman's rho (**F**, *P*-value = 0.71, *t*-test of method difference in a multiple linear regression) in low-input datasets with DNA input ranging from single cell to 250 ng. (**G**) Comparisons of three probe masking methods: (top) the total number of readings surviving detection masking from 53 FFPE samples, and (bottom) the methylation reading deviation from putative ground truths in these 53 samples.

background-induced artificial readings while minimally removing probes with biological signals (Figure 5B). In the cell line experiments, probes that survive ELBAR masking maintain a bimodal distribution of beta values as biologically expected. In contrast, pOOBAH, a prior method designed for high-input datasets, masked probes more aggressively. The probes surviving pOOBAH masking are slightly asymmetric in the beta-value envelope and show a small amount of background-dominated beta values (Figure 5B). ELBAR effectively masked these beta values associated with low signal intensity and artifactually fixed around 0.5.

Testing ELBAR on public EPIC, HM450 (Figure 5C, Supplementary Table S2A), and MM285 (Figure 5D) datasets, we found that it could rescue a significant number of probes compared to pOOBAH. Interestingly, experiments with array-wide failure remain low in detection rates, suggesting ELBAR can discriminate probe failure against array-wide failures. The probes rescued by ELBAR from pOOBAH show biological relevance, evidenced by higher correlation with the 250 ng datasets (Supplementary Figure S5B). Of note, ELBAR combines negative control probes for background calibration and only considers intermediate methylation reading from low-intensity probes. Hence, its masking would not be influenced by true biological methylations. For example, we validated ELBAR's performance in samples with globally high, intermediate, and low methylation levels (Supplementary Figures S5C–E), including testicular seminoma tissues (Supplementary Figure S5F). ELBAR improves detection in wide input ranges (Figure 5E, *P*-value = 2.7E-8, *t*-test of the method sensitivity difference in multiple linear regression) without harming accuracy (Figure 5F, no statistical significance detected from method accuracy differences).

To further validate ELBAR's performance in FFPE samples, we compared ELBAR, pOOBAH and minfi's detectionP function in a previous study of 53 melanoma FFPE tissue samples (53). We used five samples with the best detection *p*-values to derive a putative ground truth methylation profile and evaluated the measurement deviation in probe sets stratified by the masking status under the three methods (Figure 5G). Probes that survive all three masking methods have the lowest methylation level deviation as expected, followed by the probes masked only by pOOBAH. These probes are the greatest in number compared to other probe masking groups, suggesting that pOOBAH may have caused a significant loss of biological signal in this dataset. Overall, ELBAR-masked probes are associated with higher measurement deviation from the ground truth (dark red in Figure 5G), despite that they may survive the masking by the other two methods. Collectively, these results point to an advantage of using ELBAR for detection calling over pOOBAH and detectionP in FFPE samples.

Low-input BeadChip data captures the demethylation dynamics in primordial germ cells

PGCs are typically present in low numbers, hindering their analysis by the standard Infinium array processing workflow (54–56). In mammals, PGCs undergo genome-wide epigenetic reprogramming, including global DNA methylation loss, as they migrate from the epiblast to the bipotential gonads (57). This corresponds to embryonic day(E)7.5 to E14.5 of development in the mouse. We applied our optimized method to study the methylation of mouse gonadal PGCs collected at E11.5 to E14.5 (Figure 6A). For each time point, PGCs

from a pair of embryonic gonads were FACS sorted (Methods), and the aliquoted volume varied from 0.25 µl to 9 µl. We employed workflow J, with pre-amplified DNA amounts ranging from ~1 ng to 13 ng (Supplementary Table S2B). As a contrast, we included methylome profiles of mouse liver, lung, ovary, and testes tissues in our analysis (Methods, Supplementary Table S2C).

Consistent with prior knowledge, PGCs exhibit the lowest methylation level relative to somatic tissues and adult gonads across major chromatin states (Figure 6B and C). Consistent with the probe design rationale, 'PGCMeth' probes showed resistance to methylation in E11.5 and E13.5 PGCs. The genome-wide distribution of PGC methylation loss is largely uncoupled from their methylation states observed in non-PGC tissues. For example, the genomic regions with active gene transcription (ChromHMM states 'Tx' and 'TxWk') were associated with the highest methylation in non-PGC tissues (Figure 6B). But in PGCs, gene bodies are less methylated than heterochromatin and transcriptionally quiescent regions (Figure 6B and C). We observed a partial methylation loss in E11.5 PGCs. The demethylation process culminated in E13.5 PGCs. Male PGCs initiated re-methylation as early as E14.5, while female PGCs remained similarly unmethylated at E14.5 as E13.5. This is consistent with *de novo* methylation occurring in female germline postnatally as oocytes are recruited for growth during each reproductive cycle (58). This sex disparity in methylation rebound is evident in imprinted gene-associated differentially methylated regions (DMRs) and gene bodies (Figure 6D and Supplementary Figure S6A).

The arrays allow for detailed analysis of the timing of the DNA methylation change across genomic elements. For example, the retained DNA methylation at heterochromatin is enriched at TRIM28 binding sites (Figure 6E and Supplementary Figure S6B). TRIM28 regulates the transcription of TE, particularly endogenous retroviruses (ERV) (59). Specific ERV elements are known to retain DNA methylation in human PGC development (60) and mice (61). These observations suggest a critical role for DNA methylation in TE suppression for maintaining germ cell genome integrity and intergenerational epigenetic inheritance.

Interestingly, PGC residual methylation is also enriched for the binding of the DNA methylation reader proteins MECP2 and MBD1, supporting their reported roles in DNA methylation-mediated TE regulation (62). Among TE DNA families, LTR elements were more resistant to PGC demethylation than SINEs and LINEs, consistent with their functions as transcriptional promoters (63). The role of methylation retention in TE regulation is further supported by the higher methylation retention in evolutionarily younger TE subfamilies than older TE families (64) (Figure 6F). Intracisternal A Particle (IAP) elements were previously highlighted to exhibit extensive methylation retention in PGCs (65). We identified a heterogeneous pattern of their demethylation dynamics in PGCs. IAPLTR2, IAPEY and IAPA are among the most resistant families, while IAPEY5, IAPLTR4 and IAP5 are the least methylated.

In addition to TEs, imprinted gene DMRs, germ cells, and placenta-specific hypomethylated sites show later DNA methylation loss compared to the rest of the genome. This is due to active DNA demethylation pathways, mediated by TET proteins, that are required for methylation erasure at imprinting control regions (39,66). CGs flanked by A/Ts are more susceptible to aging-associated DNA demethylation (51). We
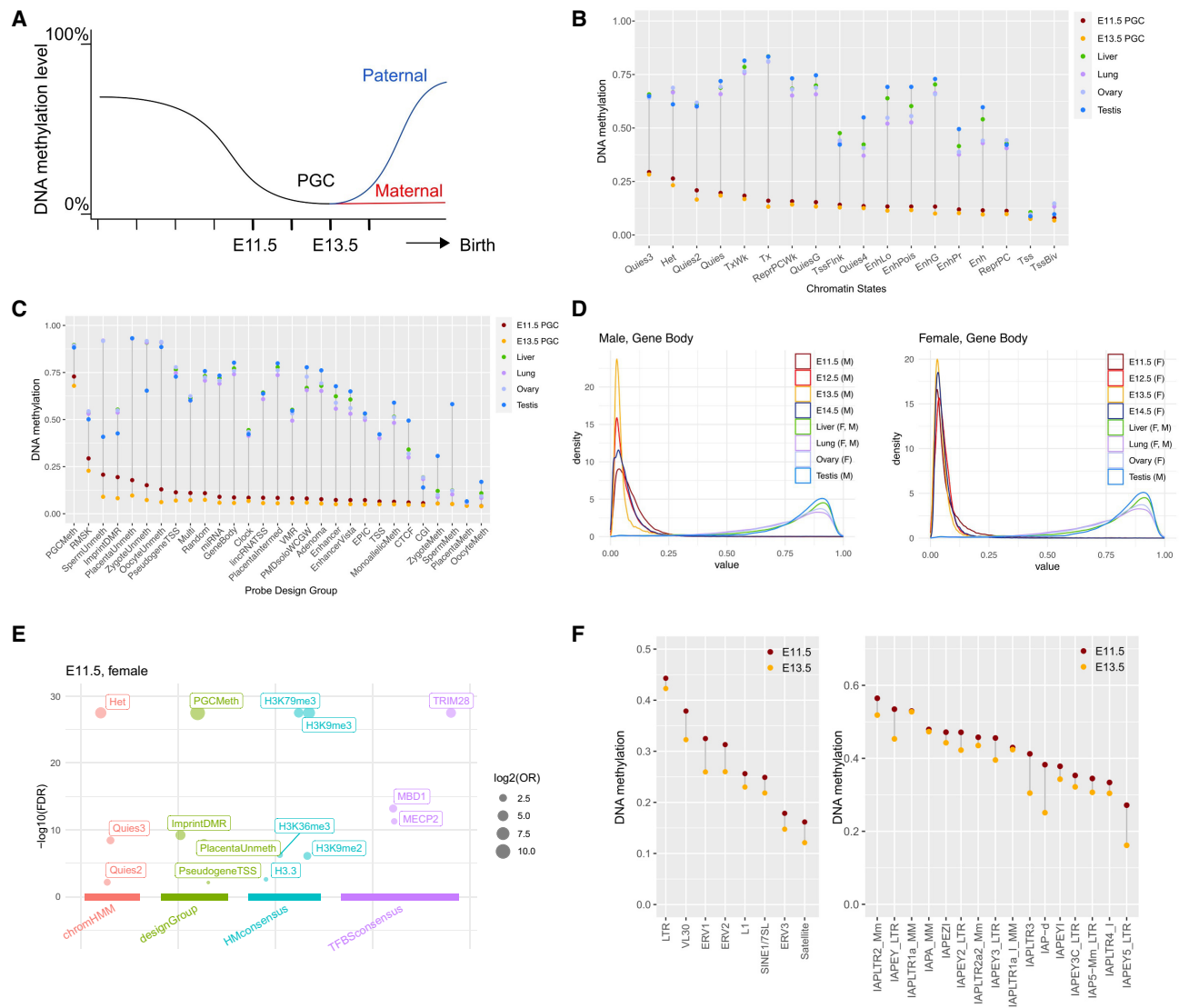
Figure 6. Low-input Infinium array data reveals epigenetic dynamics of primordial germ cell development. (**A**) Expected DNA methylation dynamics during primordial germ cell developments. E11.5 and E13.5 refer to embryonic days after fertilization. (B, C) DNA demethylation patterns in PGCs and tissues stratified by ChromHMM states (**B**) and mouse array probe design groups (**C**). (**D**) DNA methylation distribution in gene body regions in PGCs and tissues. M: male; F: female. (**E**) Enriched genomic features by CpGs that retain methylation level (over 0.5) in the three female E11.5 PGCs. The size of each circle represents the log2 odds ratio (OR) of individual curated CpG sets, and the y-axis indicates the -log10 of the false discovery rate (FDR) for each curated CpG set. Arranged from left to right along the x-axis are four distinct sets of databases: chromatin states, probe design group, histone modification consensus, and transcription factor binding site consensus motif. (**F**) DNA demethylation patterns in TEs (left) and IAP elements (right) across PGCs and tissues.

did not observe this sequence context preference for PGC development ([Supplementary Figure S6](#)C), suggesting a distinct, TET-mediated demethylation mechanism.

Discussion

Despite the successful employment of Infinium BeadChips in population-scale DNA methylome studies, their potential for 'difficult' DNA, i.e. when the input is limited in quality, quantity, or both, has not been fully explored or optimized. This restricts the scope of the Infinium array usage for cell-free DNA, microdissected tissues, and other samples of limited availability. Here, we presented experimental and computational resources to enable array usage in these suboptimal settings, especially with low-input DNA. Experimentally, we explored the array's compatibility with random priming-based whole genome preamplification and enzymatic base conversion by TET/APOBEC3A. Our data suggests that both preamplification and enzymatic base conversion by TET/APOBEC3A using the NEBNext Enzymatic Methyl-seq (Workflow J). Computationally, we developed ELBAR for preserving biological signals from suboptimal input datasets. ELBAR excludes only probes dominated by background noises.

Besides, we comprehensively characterized the biological and technical determinants of array performance from public datasets. From surveying 100 000+ datasets and using probe detection rate as the main performance metric, we found that cell-free DNA, saliva, bone and FFPE samples, tend to have worse detection rates compared to cultured cells, primary and fresh frozen tissues. FFPE samples and those of sheer lower input show a different genomic distribution in signal detection.

Plasma cfDNA tends to cause detection loss at CpG-sparse, GC-low genomic regions and preserved detection at CpG-rich regions such as bivalent chromatin. In contrast, FFPE-induced detection loss is less biased across genomic regions. This low-input sample bias can be attributed to the weaker intrinsic signal from GC-low probes (Figure 4C and D). As cfDNA localization is known to be linked to nucleosome footprints and can inform cell of origin (67), the array signal intensity bias may serve as an unconventional source of cell of origin signal to complement the methylation signal that the array data already carries.

Although Illumina recommends a minimum of 200 ng of DNA for current Infinium BeadChips, there has been research exploring lower input amounts. For example, reproducible results were achieved with over 125 ng of DNA from peripheral blood (68), and an input of 16 ng exhibited a high correlation with 500 ng input, albeit with lower reproducibility. Another study identified 75 ng as the minimum requirement (69), and a few other studies suggested that 10 ng is acceptable (35,70,71). The fact that the standard protocol works at these sub-optimal input ranges is likely due to the isothermal amplification. However, the true lower input limit of the technology remains underexplored, and the extent to which precision and sensitivity are compromised with decreased DNA input remains incompletely resolved. Our work shows that Infinium arrays are reasonably compatible with picogram-range input DNA or single-digit cell number. Preamplification using Klenow fragments and enzymatic conversion further magnifies probe signal intensities, leading to reproducible profiling of five-cell methylomes. Interestingly, additional adapter-based PCR amplification did not lead to a further increase in probe detection or accuracy (Supplementary Table S1D), likely due to loss of library complexity from amplification bias (72).

The Infinium technology is currently not cost-competitive for projects that only massively profile methylomes in single cells due to the incompatibility with the sample multiplexing (73). The focus of this work is to explore the capacity of Infinium technology in profiling samples of variable quantities and whether the data is comparable. This is most relevant to applications where DNA can be of high or limited quantities, such as microdissected tissues (24) and cell-free DNA. In fact, techniques shared with sequencing-based methods, such as preamplification using Klenow or other polymerases (27), would benefit both high and low input samples (e.g. the 50-ng samples in Figure 2C and D). We showed that low-input Infinium BeadChip data is comparable to high-input data and is biologically relevant. It reflects the allelic nature of the DNA methylation signal on low DNA input. Intermediate methylation levels were resolved to high and low methylation readings. These binary readings reflect cell-to-cell heterogeneity and when merged, their population averages recapitulated bulk input measurements. Our single-cell array data reached 20% detection, similar to previous deep single-cell WGBS datasets (30).

Previous data analysis workflows relied on a single threshold for determining signal detection success (74). While this is a viable assumption for the high input data, it does not always hold for low-input datasets. In low-input datasets, biological signals overlap more with background signals in signal intensity, particularly for probes with intrinsically low foreground signals. Our analysis showed that this intrinsic propensity is linked to the number of Cs in the probe sequences, likely rem-iniscent of a GC content bias as the probes are designed to be G-less to pair with converted genomic DNA. The stronger overlaps of biological with background signal not only obfuscate the detection discrimination but also bias the beta value calculation towards 0.5 due to the incomplete subtraction of signal background from the observed compound signal. Users should consider this major tradeoff of measurement precision for sensitivity.

Compared to pOOBAH and other detection calling methods designed to minimize false discovery in high-input datasets, ELBAR seeks to mask only probes fully dominated by signal background, leaving probes with true biological signals visible in downstream analysis. However, we caution that probe readings surviving ELBAR detection may be influenced by background signals to various degrees, leading to unstable quantitative accuracy. The detection *p*-values can serve as measures of background influences besides their traditional use for probe masking. For high-input samples, pOOBAH and ELBAR perform similarly (Figure 5E and F), suggesting that one may use ELBAR for data from all input settings.

Despite the reduced probe detection in low-input datasets, probes that target multi-copy DNA, such as mitochondria and repetitive elements (e.g. the B1 elements and satellite sequences), retain high signal intensities. In the low-input datasets, we observed that these probes measure aggregated signals from multiple genomic loci, making an unconventional use of the methylation BeadChips—as a tool to study the global epigenetic regulation of multi-copy TEs. In our work, we applied our low-input protocol to profile mouse PGCs. We validated the dynamics of global methylation erasure in PGCs, a sex disparity in remethylation, as well as the demethylation resistance at TRIM28 binding sites which are known to escape germ cell epigenetic remodeling (60). These multi-mapping probes also revealed that evolutionarily younger LTR repeat families retained more methylation than other repeat families. These methylation retentions can protect germ-line genome integrity from TE transcriptional mobilization.

Conclusion

We presented experimental and computational solutions for applying Infinium BeadChips to low-input and single-cell samples. Based on whole-genome preamplification and enzymatic base conversion, our new methods revealed a previously underappreciated low-input potential of this popular methylation profiling assay. We demonstrated the power of these methods by applying them to uncover detailed demethylation dynamics of murine primordial germ cell development.

Data availability

All BeadChip data produced in this study is available through GEO (accession: GSE239290). ELBAR and other informatics for low input methylation BeadChip are implemented in the SeSAMe (version 1.18.4+) available through Bioconductor (https://doi.org/doi:10.18129/B9.bioc.sesame). ELBAR can also be used in the openSesame workflow with the "I" code specified in the prep= argument (see SeSAMe vignette).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Funding

Conflict of interest statement

W.Z. received Infinium BeadChips from Illumina Inc.

References

1. Greenberg,M.V.C. and Bourc'his,D. (2019) The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.*, **20**, 590–607.
2. Bibikova,M., Barnes,B., Tsan,C., Ho,V., Klotzle,B., Le,J.M., Delano,D., Zhang,L., Schroth,G.P., Gunderson,K.L., *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.
3. Maden,S.K., Thompson,R.F., Hansen,K.D. and Nellore,A. (2021) Human methylome variation across Infinium 450K data on the Gene Expression Omnibus. *NAR Genom. Bioinform.*, **3**, lqab025.
4. Hawe,J.S., Wilson,R., Schmid,K.T., Zhou,L., Lakshmanan,L.N., Lehne,B.C., Kuhnel,B., Scott,W.R., Wielscher,M., Yew,Y.W., *et al.* (2022) Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat. Genet.*, **54**, 18–29.
5. Min,J.L., Hemani,G., Hannon,E., Dekkers,K.F., Castillo-Fernandez,J., Luijk,R., Carnero-Montoro,E., Lawson,D.J., Burrows,K., Suderman,M., *et al.* (2021) Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.*, **53**, 1311–1321.
6. Thompson,M., Hill,B.L., Rakocz,N., Chiang,J.N., Geschwind,D., Sankararaman,S., Hofer,I., Cannesson,M., Zaitlen,N. and Halperin,E. (2022) Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *NPJ Genom Med*, **7**, 50.
7. Aref-Eshghi,E., Kerkhof,J., Pedro,V.P., Groupe,D.I.F., Barat-Houari,M., Ruiz-Pallares,N., Andrau,J.C., Lacombe,D., Van-Gils,J., Fergelot,P., *et al.* (2020) Evaluation of DNA methylation episignatures for diagnosis and phenotype correlations in 42 mendelian neurodevelopmental disorders. *Am. J. Hum. Genet.*, **106**, 356–370.
8. Li,M., Zou,D., Li,Z., Gao,R., Sang,J., Zhang,Y., Li,R., Xia,L., Zhang,T., Niu,G., *et al.* (2019) EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic. Acids. Res.*, **47**, D983–D988.
9. Battram,T., Yousefi,P., Crawford,G., Prince,C., Sheikhali Babaei,M., Sharp,G., Hatcher,C., Vega-Salas,M.J., Khodabakhsh,S., Whitehurst,O., *et al.* (2022) The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res*, **7**, 41.
10. Berdasco,M. and Esteller,M. (2019) Clinical epigenetics: seizing opportunities for translation. *Nat. Rev. Genet.*, **20**, 109–127.
11. Capper,D., Jones,D.T.W., Sill,M., Hovestadt,V., Schrimpf,D., Sturm,D., Koelsche,C., Sahm,F., Chavez,L., Reuss,D.E., *et al.* (2018) DNA methylation-based classification of central nervous system tumours. *Nature*, **555**, 469–474.
12. Li,H.T., Xu,L., Weisenberger,D.J., Li,M., Zhou,W., Peng,C.C., Stachelek,K., Cobrinik,D., Liang,G. and Berry,J.L. (2022) Characterizing DNA methylation signatures of retinoblastoma using aqueous humor liquid biopsy. *Nat. Commun.*, **13**, 5523.
13. Mannens,M., Lombardi,M.P., Alders,M., Henneman,P. and Bliek,J. (2022) Further introduction of DNA methylation (DNAm) arrays in regular diagnostics. *Front. Genet.*, **13**, 831452.
14. Haghani,A., Li,C.Z., Robeck,T.R., Zhang,J., Lu,A.T., Ablaeva,J., Acosta-Rodriguez,V.A., Adams,D.M., Alagaili,A.N., Almunia,J., *et al.* (2023) DNA methylation networks underlying mammalian traits. *Science*, **381**, eabq5693.
15. Ding,W., Kaur,D., Horvath,S. and Zhou,W. (2023) Comparative epigenome analysis using Infinium DNA methylation BeadChips. *Brief. Bioinform.*, **24**, bbac617.
16. Arneson,A., Haghani,A., Thompson,M.J., Pellegrini,M., Kwon,S.B., Vu,H., Maciejewski,E., Yao,M., Li,C.Z., Lu,A.T., *et al.* (2022) A mammalian methylation array for profiling methylation levels at conserved sequences. *Nat. Commun.*, **13**, 783.
17. Moratilla,A., Sainz de la Maza,D., Cadenas Martin,M., Lopez-Iglesias,P., Gonzalez-Peramato,P. and De Miguel,M.P. (2021) Inhibition of PKCepsilon induces primordial germ cell reprogramming into pluripotency by HIF1&2 upregulation and histone acetylation. *Am. J. Stem Cells*, **10**, 1–17.
18. Vasioukhin,V., Anker,P., Maurice,P., Lyautey,J., Lederrey,C. and Stroun,M. (1994) Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.*, **86**, 774–779.
19. Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nat. Rev. Cancer*, **3**, 253–266.
20. Lenaerts,L., Tuveri,S., Jatsenko,T., Amant,F. and Vermeesch,J.R. (2020) Detection of incipient tumours by screening of circulating plasma DNA: hype or hope? *Acta Clin. Belg.*, **75**, 9–18.
21. Vidaki,A. and Kayser,M. (2017) From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. *Genome Biol.*, **18**, 238.
22. Moss,J., Magenheim,J., Neiman,D., Zemmour,H., Loyfer,N., Korach,A., Samet,Y., Maoz,M., Druid,H., Arner,P., *et al.* (2018) Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.*, **9**, 5068.
23. Houseman,E.A., Molitor,J. and Marsit,C.J. (2014) Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, **30**, 1431–1439.
24. Gagnon,J.F., Sanschagrin,F., Jacob,S., Tremblay,A.A., Provencher,L., Robert,J., Morin,C. and Diorio,C. (2010) Quantitative DNA methylation analysis of laser capture microdissected formalin-fixed and paraffin-embedded tissues. *Exp. Mol. Pathol.*, **88**, 184–189.
25. Siegmund,K.D., Marjoram,P., Woo,Y.J., Tavare,S. and Shibata,D. (2009) Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 4828–4833.
26. Hernandez,H.G., Sandoval-Hernandez,A.G., Garrido-Gil,P., Labandeira-Garcia,J.L., Zelaya,M.V., Bayon,G.F., Fernandez,A.F., Fraga,M.F., Arboleda,G. and Arboleda,H. (2018) Alzheimer's disease DNA methylome of pyramidal layers in frontal cortex: laser-assisted microdissection study. *Epigenomics*, **10**, 1365–1382.
27. Aaltonen,K.E., Ebbesson,A., Wigerup,C. and Hedenfalk,I. (2011) Laser capture microdissection (LCM) and whole genome amplification (WGA) of DNA from normal breast tissue — optimization for genome wide array analyses. *BMC Res. Notes*, **4**, 69.
28. Ahn,J., Heo,S., Lee,J. and Bang,D. (2021) Introduction to single-cell DNA methylation profiling methods. *Biomolecules*, **11**, 1013.
29. Iqbal,W. and Zhou,W. (2023) Computational methods for single-cell DNA methylome analysis. *Genomics Proteomics Bioinformatics*, **21**, 48–66.
30. 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.

31. Hui,T., Cao,Q., Wegrzyn-Woltosz,J., O'Neill,K., Hammond,C.A., Knapp,D., Laks,E., Moksa,M., Aparicio,S., Eaves,C.J., *et al.* (2018) High-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. *Stem Cell Rep.*, **11**, 578–592.

32. Wang,T., Loo,C.E. and Kohli,R.M. (2022) Enzymatic approaches for profiling cytosine methylation and hydroxymethylation. *Mol Metab*, **57**, 101314.

33. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

34. Zhou,W., Triche,T.J. Jr, Laird,P.W. and Shen,H. (2018) SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.*, **46**, e123.

35. Whalley,C., Payne,K., Domingo,E., Blake,A., Richman,S., Brooks,J., Batis,N., Spruce,R., Consortium,S.C., Mehanna,H., *et al.* (2021) Ultra-low DNA input into whole genome methylation assays and detection of oncogenic methylation and copy number variants in circulating tumour DNA. *Epigenomes*, **5**, 6.

36. Moran,S., Vizoso,M., Martinez-Cardus,A., Gomez,A., Matias-Guiu,X., Chiavenna,S.M., Fernandez,A.G. and Esteller,M. (2014) Validation of DNA methylation profiling in formalin-fixed paraffin-embedded samples using the Infinium HumanMethylation450 Microarray. *Epigenetics*, **9**, 829–833.

37. Ghantous,A., Saffery,R., Cros,M.P., Ponsonby,A.L., Hirschfeld,S., Kasten,C., Dwyer,T., Herceg,Z. and Hernandez-Vargas,H. (2014) Optimized DNA extraction from neonatal dried blood spots: application in methylome profiling. *BMC Biotechnol.*, **14**, 60.

38. Lengner,C.J., Camargo,F.D., Hochedlinger,K., Welstead,G.G., Zaidi,S., Gokhale,S., Scholer,H.R., Tomilin,A. and Jaenisch,R. (2007) Oct4 expression is not required for mouse somatic stem cell self-renewal. *Cell Stem Cell*, **1**, 403–415.

39. Prasasya,R.D., Caldwell,B.A., Liu,Z., Wu,S., Leu,N.A., Fowler,J.M., Cincotta,S.A., Laird,D.J., Kohli,R.M. and Bartolomei,M.S. (2023) TET1 Catalytic activity is required for reprogramming of imprinting control regions and patterning of sperm-specific hypomethylated regions. bioRxiv doi: https://doi.org/10.1101/2023.02.21.529426, 21 February 2023, preprint: not peer reviewed.

40. Gross,J.A., Lefebvre,F., Lutz,P.E., Bacot,F., Vincent,D., Bourque,G. and Turecki,G. (2016) Variations in 5-methylcytosine and 5-hydroxymethylcytosine among human brain, blood, and saliva using oxBS and the Infinium MethylationEPIC array. *Biol. Methods Protoc*, **1**, 1–8.

41. Solomon,O., Macisaac,J.L., Tindula,G., Kobor,M.S., Eskenazi,B. and Holland,N. (2019) 5-Hydroxymethylcytosine in cord blood and associations of DNA methylation with sex in newborns. *Mutagenesis*, **34**, 315–322.

42. Grit,J.L., Johnson,B.K., Dischinger,P.S., C,J.E., Adams,M., Campbell,S., Pollard,K., Pratilas,C.A., Triche,T.J. Jr, Graveel,C.R, *et al.* (2021) Distinctive epigenomic alterations in NF1-deficient cutaneous and plexiform neurofibromas drive differential MKK/p38 signaling. *Epigenetics Chromatin*, **14**, 7.

43. Zhou,W., Hinoue,T., Barnes,B., Mitchell,O., Iqbal,W., Lee,S.M., Foy,K.K., Lee,K.H., Moyer,E.J., VanderArk,A., *et al.* (2022) DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. *Cell Genom*, **2**, 100144–100153.

44. Hill,P.W.S., Leitch,H.G., Requena,C.E., Sun,Z., Amouroux,R., Roman-Trufero,M., Borkowska,M., Terragni,J., Vaisvila,R., Linnett,S., *et al.* (2018) Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature*, **555**, 392–396.

45. Shen,H., Shih,J., Hollern,D.P., Wang,L., Bowlby,R., Tickoo,S.K., Thorsson,V., Mungall,A.J., Newton,Y., Hegde,A.M., *et al.* (2018)

46. Integrated molecular characterization of testicular germ cell tumors. *Cell Rep.*, **23**, 3392–3406.

46. Gao,G.F., Parker,J.S., Reynolds,S.M., Silva,T.C., Wang,L.B., Zhou,W., Akbani,R., Bailey,M., Balu,S., Berman,B.P., *et al.* (2019) Before and after: comparison of legacy and harmonized TCGA Genomic Data Commons' Data. *Cell Syst.*, **9**, 24–34.

47. Bibikova,M., Le,J., Barnes,B., Saedinia-Melnyk,S., Zhou,L., Shen,R. and Gunderson,K.L. (2009) Genome-wide DNA methylation profiling using Infinium(R) assay. *Epigenomics*, **1**, 177–200.

48. Wang,L., Zhang,J., Duan,J., Gao,X., Zhu,W., Lu,X., Yang,L., Zhang,J., Li,G., Ci,W., *et al.* (2014) Programming and inheritance of parental DNA methylomes in mammals. *Cell*, **157**, 979–991.

49. Kaur,D., Lee,S.M., Goldberg,D., Spix,N.J., Hinoue,T., Li,H.-T., Dwaraka,V.B., Smith,R., Shen,H. and Liang,G. (2023) Comprehensive evaluation of the Infinium human MethylationEPIC v2 BeadChip. *Epigenetics Commun.*, **3**, 6.

50. Wong,K.K., Stillwell,L.C., Dockery,C.A. and Saffer,J.D. (1996) Use of tagged random hexamer amplification (TRHA) to clone and sequence minute quantities of DNA–application to a 180 kb plasmid isolated from Sphingomonas F199. *Nucleic Acids Res.*, **24**, 3778–3783.

51. Zhou,W., Dinh,H.Q., Ramjan,Z., Weisenberger,D.J., Nicolet,C.M., Shen,H., Laird,P.W. and Berman,B.P. (2018) DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.*, **50**, 591–602.

52. Rooney,J.P., Ryde,I.T., Sanders,L.H., Howlett,E.H., Colton,M.D., Germ,K.E., Mayer,G.D., Greenamyre,J.T. and Meyer,J.N. (2015) PCR based determination of mitochondrial DNA copy number in multiple species. *Methods Mol. Biol.*, **1241**, 23–38.

53. Pradhan,D., Jour,G., Milton,D., Vasudevaraja,V., Tetzlaff,M.T., Nagarajan,P., Curry,J.L., Ivan,D., Long,L., Ding,Y., *et al.* (2019) Aberrant DNA methylation predicts melanoma-specific survival in patients with acral melanoma. *Cancers (Basel)*, **11**, 2031.

54. Leitch,H.G., Nichols,J., Humphreys,P., Mulas,C., Martello,G., Lee,C., Jones,K., Surani,M.A. and Smith,A. (2013) Rebuilding pluripotency from primordial germ cells. *Stem Cell Rep.*, **1**, 66–78.

55. Teng,H., Sui,X., Zhou,C., Shen,C., Yang,Y., Zhang,P., Guo,X. and Huo,R. (2016) Fatty acid degradation plays an essential role in proliferation of mouse female primordial germ cells via the p53-dependent cell cycle regulation. *Cell Cycle*, **15**, 425–431.

56. Saitou,M. and Yamaji,M. (2012) Primordial germ cells in mice. *Cold Spring Harb. Perspect. Biol.*, **4**, a008375.

57. Kurimoto,K. and Saitou,M. (2019) Germ cell reprogramming. *Curr. Top. Dev. Biol.*, **135**, 91–125.

58. Gkountela,S., Zhang,K.X., Shafiq,T.A., Liao,W.W., Hargan-Calvopina,J., Chen,P.Y. and Clark,A.T. (2015) DNA demethylation dynamics in the Human prenatal germline. *Cell*, **161**, 1425–1436.

59. Rowe,H.M., Jakobsson,J., Mesnard,D., Rougemont,J., Reynard,S., Aktas,T., Maillard,P.V., Layard-Liesching,H., Verp,S., Marquis,J., *et al.* (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, **463**, 237–240.

60. Tang,W.W., Dietmann,S., Irie,N., Leitch,H.G., Floros,V.I., Bradshaw,C.R., Hackett,J.A., Chinnery,P.F. and Surani,M.A. (2015) A unique gene regulatory network resets the Human germline epigenome for development. *Cell*, **161**, 1453–1467.

61. Lane,N., Dean,W., Erhardt,S., Hajkova,P., Surani,A., Walter,J. and Reik,W. (2003) Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis*, **35**, 88–93.

62. Yu,F., Zingler,N., Schumann,G. and Stratling,W.H. (2001) Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res.*, **29**, 4493–4501.

63. Klaver,B. and Berkhout,B. (1994) Comparison of 5' and 3' long terminal repeat promoter function in human immunodeficiency virus. *J. Virol.*, **68**, 3830–3840.

64. Zhou,W., Liang,G., Molloy,P.L. and Jones,P.A. (2020) DNA methylation enables transposable element-driven genome expansion. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 19359–19366.

65. Seisenberger,S., Andrews,S., Krueger,F., Arand,J., Walter,J., Santos,F., Popp,C., Thienpont,B., Dean,W. and Reik,W. (2012) The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell*, **48**, 849–862.

66. Yamaguchi,S., Hong,K., Liu,R., Inoue,A., Shen,L., Zhang,K. and Zhang,Y. (2013) Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. *Cell Res.*, **23**, 329–339.

67. Snyder,M.W., Kircher,M., Hill,A.J., Daza,R.M. and Shendure,J. (2016) Cell-free DNA comprises an In vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, **164**, 57–68.

68. Christiansen,S.N., Andersen,J.D., Kampmann,M.L., Liu,J., Andersen,M.M., Tfelt-Hansen,J. and Morling,N. (2022) Reproducibility of the Infinium methylationEPIC BeadChip assay using low DNA amounts. *Epigenetics*, **17**, 1636–1645.

69. Abbasi,M., Smith,A.D., Swaminathan,H., Sangngern,P., Douglas,A., Horsager,A., Carrell,D.T. and Uren,P.J. (2018) Establishing a stable, repeatable platform for measuring changes in sperm DNA methylation. *Clin Epigenetics*, **10**, 119.

70. Watkins,S.H., Ho,K., Testa,C., Falk,L., Soule,P., Nguyen,L.V., FitzGibbon,S., Slack,C., Chen,J.T., Davey Smith,G., *et al.* (2022) The impact of low input DNA on the reliability of DNA methylation as measured by the Illumina Infinium MethylationEPIC BeadChip. *Epigenetics*, **17**, 2366–2376.

71. Hovestadt,V., Remke,M., Kool,M., Pietsch,T., Northcott,P.A., Fischer,R., Cavalli,F.M., Ramaswamy,V., Zapatka,M., Reifenberger,G., *et al.* (2013) Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol.*, **125**, 913–916.

72. Polz,M.F. and Cavanaugh,C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microb.*, **64**, 3724–3730.

73. 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.

74. Lehne,B., Drong,A.W., Loh,M., Zhang,W., Scott,W.R., Tan,S.T., Afzal,U., Scott,J., Jarvelin,M.R., Elliott,P., *et al.* (2015) A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.*, **16**, 37.

75. van der Velde,A., Fan,K., Tsuji,J., Moore,J.E., Purcaro,M.J., Pratt,H.E. and Weng,Z. (2021) Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun. Biol.*, **4**, 239.