Sensitive and reproducible cell-free methylome quantification with synthetic spike-in controls

Samantha L. Wilson (0000-0003-4346-9696)[1,*], Shu Yi Shen (0000-0001-7036-2291)[1,*],
Lauren Harmon (0000-0002-2248-0602)[2], Justin M. Burgener (0000-0001-9522-8865)[1,3],
Tim Triche Jr. (0000-0001-5665-946X)[2], Scott V. Bratman (0000-0001-8610-4908)[1,3],
Daniel D. De Carvalho (0000-0002-8572-5259)[1,3], and Michael M. Hoffman (0000-0002-4517-1562)[1,3,4,5]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
[2]Van Andel Institute, Grand Rapids, MI, United States of America
[3]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[4]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[5]Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[*]Authors contributed equally to this work

February 12, 2021

Abstract

Background. Cell-free methylated DNA immunoprecipitation-sequencing (cfMeDIP-seq) identifies genomic regions with DNA methylation, using a protocol adapted to work with low-input DNA samples and with cell-free DNA (cfDNA). This method allows for DNA methylation profiling of circulating tumour DNA in cancer patients' blood samples. Such epigenetic profiling of circulating tumour DNA provides information about in which tissues tumour DNA originates, a key requirement of any test for early cancer detection. In addition, DNA methylation signatures provide prognostic information and can detect relapse. For robust quantitative comparisons between samples, immunoprecipitation enrichment methods like cfMeDIP-seq require normalization against common reference controls.

Methods. To provide a simple and inexpensive reference for quantitative normalization, we developed a set of synthetic spike-in DNA controls for cfMeDIP-seq. These controls account for technical variation in enrichment efficiency due to biophysical properties of DNA fragments. Specifically, we designed 54 DNA fragments with combinations of methylation status (methylated and unmethylated), fragment length (80 bp, 160 bp, 320 bp), G+C content (35%, 50%, 65%), and fraction of CpG dinucleotides within the fragment (1/80 bp, 1/40 bp, 1/20 bp). We ensured that the spike-in synthetic DNA sequences do not align to the human genome. We integrated unique molecular indices (UMIs) into cfMeDIP-seq to control for differential amplification after enrichment. To assess enrichment bias according to distinct biophysical properties, we conducted cfMeDIP-seq solely on spike-in DNA fragments. To optimize the amount of spike-in DNA required, we added varying quantities of spike-in control DNA to sheared HCT116 colon cancer genomic DNA prior to cfMeDIP-seq. To assess batch effects, three separate labs conducted cfMeDIP-seq on peripheral blood plasma samples from acute myeloid leukemia (AML) patients.

Results. We show that cfMeDIP-seq enriches for highly methylated regions, capturing ≥99.99% of methylated spike-in control fragments with ≤0.01% non-specific binding and preference for both high G+C content fragments and fragments with more CpGs. The use of 0.01 ng of spike-in control DNA total provided sufficient sequencing reads to adjust for variance due to fragment length, G+C content, and CpG fraction. Using the known amount of each spiked-in fragment, we created a generalized linear model that absolutely quantifies molar amount from read counts across the genome, while adjusting for fragment length, G+C content, and CpG fraction. Employing our spike-in controls greatly mitigates batch effects, reducing batch-associated variance to ≤ 1% of the total variance within the data.

Discussion. Incorporation of spike-in controls enables absolute quantification of methylated cfDNA generated from methylated DNA immunoprecipitation-sequencing (MeDIP-seq) experiments. It mitigates batch effects and corrects for biases in enrichment due to known biophysical properties of DNA fragments and other technical biases. We created an R package, `spiky`, to convert read counts to picomoles of DNA fragments, while adjusting for fragment properties that affect enrichment. The `spiky` package is available on GitHub (`https://github.com/trichelab/spiky`) and will soon be available on Bioconductor.

Contact: michael.hoffman@utoronto.ca

1 Introduction

Cell-free methylated DNA immunoprecipitation-sequencing (cfMeDIP-seq) identifies DNA methylation using low-input samples of cell-free DNA (cfDNA). This method detects DNA methylation patterns reflective of distinct cancer types from circulating tumour DNA, which arises from tumour cells shedding DNA into an individual's blood.[1,2] cfMeDIP-seq proves ideal when assessing peripheral blood plasma from cancer patients, where one may obtain only a small amount of circulating tumour DNA and no indication of from where the circulating tumour DNA originates.

Sequencing assay methods, such as cfMeDIP-seq or RNA-seq, require a control for more accurate quantitative comparison across samples and batches. Reference controls for sequencing assays have consisted of spike-in reference DNA or RNA fragments of known sequence.[3–7] Without spike-in controls, one must assume that a given amount of assayed material produces equal DNA or RNA yields in different experimental conditions, and that this also holds true across all genomic regions.[4] By normalizing quantification to a known amount of spike-in DNA put into a sample, we can overcome this assumption, leading to more accurate results.[4] When carefully designed, spike-in controls can adjust for specific technical biases.

In addition to adjusting for technical biases, spike-in controls act as experimental standards for quality control. The addition of spike-in controls dramatically changes the interpretation of RNA-seq, chromatin immunoprecipitation-sequencing (ChIP-seq), and other genomic assay results.[3–7] As such, all quantitative genome-wide assays would benefit from the addition of spike-in controls.[4]

The most common approach to normalizing sequencing assay data consists in dividing the number of reads at each genomic region by the total number of reads genome-wide. This approach addresses technical variance due to sequencing depth, but it can mask differences in biological variables of interest. Normalizing data to a known amount of spike-in DNA for each sample allows for more accurate detection of differences and adjustment of biophysical properties of DNA fragments that can influence results.[4,5]

While other genomic assays have long utilized spike-in controls, methods measuring genome-wide DNA methylation have rarely used them. The previous cfMeDIP-seq protocol uses methylated and unmethylated *Arabidopsis thaliana* DNA as a spike-in control to assess immunoprecipitation and binding efficiency to methylated DNA.[2] This approach cannot correct for properties of specific methylated DNA fragments likely to influence results, such as G+C content, fragment length, and CpG fraction. Other controls used in DNA methylation enrichment methods include setting aside and sequencing a portion of input DNA without the enrichment procedure.[8] This provides a reference point to assess enrichment of DNA methylation overall. Sequencing input DNA, however, cannot adjust for properties of individual fragments that can affect enrichment. Spike-in controls for bisulfite conversion methods for assaying DNA methylation[9,10] have no use in bisulfite-free enrichment methods, like cfMeDIP.

Here, we introduce new synthetic DNA spike-in controls for cfMeDIP-seq. Our synthetic spike-ins can measure how DNA fragment properties, such as length, G+C content, and number of CpGs, can affect the number of reads produced by some known amount of fragment. Our spike-in controls assess non-specific binding, an integral part of cfMeDIP-seq analysis. The spike-in controls also mitigate technical effects such as experiments performed by different labs. We also use unique molecular indices (UMIs) to adjust for polymerase chain reaction (PCR) bias. To calculate methylation specificity, we compare methylated fragments to unmethylated fragments after cfMeDIP-seq. We add a known molar amount of spike-in controls to each sample. With this information, we apply a generalized linear model to calculate molar amount, accounting for fragment length, G+C content, and CpG fraction. These spike-in controls generate an absolute quantitative measure of methylated DNA, allowing for more robust comparisons between samples and experiments.

2 Methods

2.1 Designing synthetic DNA spike-in controls

We used public paired-end, whole genome cfDNA sequence data to assess typical cfDNA fragment properties including fragment length, G+C content, and the number of CpG dinucleotides.[11] We considered the number of CpGs as a fraction of fragment length.

From the observed distribution of cfDNA fragment properties, we set the following spike-in fragment parameters:

- 3 fragment lengths: 80 bp, 160 bp, and 320 bp
- 3 G+C contents: 35%, 50%, and 65%
- 3 CpG fractions: 1/80 bp, 1/40 bp, and 1/20 bp

These parameters generate 27 fragment combinations (3 fragment lengths × 3 G+C contents × 3 CpG fractions = 27).

We set the CpG fraction parameters so that every fragment length would have an integer number of CpGs. For example, the 80 bp fragments have 1, 2, or 4 CpG dinucleotides, and the 160 bp fragments have 2, 4, and 8 CpG dinucleotides.

We used GenRGenS version 2.0[12] to construct 27 different first-order Markov models that generate sequences with the desired parameters. For each Markov model, we generated numerous sequences. We then identified those sequences that fulfilled two criteria: (1) no alignment to human genome and (2) no potential secondary structures.

Using blastn,[13] we searched for alignment of the generated sequences to the human reference genome (GRCh38/hg38).[14] We ensured no alignment of each synthetic sequence to the genome, and selected the sequences with the lowest E-values in each search.

We used UNAFold software[15] (Integrated DNA Technologies (IDT), Coralville, IA, USA) to check for secondary DNA structure for 80 bp and 160 bp fragments. For 320 bp fragments, we used RNAstructure version 6.2[16] to check for secondary DNA structures.

We picked two sequences from those fulfilling our criteria for each of the 27 Markov models. We selected one sequence for a methylated fragment, and one for an unmethylated fragment. This produced 54 desired spike-in control sequences.

2.2 Synthetic fragment preparation

We acquired synthetic fragments for the spike-in control sequences from a commercial service (IDT, Coralville, IA, USA). For the 80 bp and 160 bp fragments we used 4 nmol Ultramer DNA Oligonucleotides. For the 320 bp fragments we used gBlocks Gene Fragments, obtaining 250 ng of each designed fragment. The two 160 bp fragments with 35% G+C content and 1/20 bp CpG fraction failed commercial design procedures, leaving us with 52 spike-in control fragments (Supplementary Table 1).

We amplified the synthetic fragments by PCR. We used the High-Fidelity 2X Master Mix (New England Biolabs, Ipswich, MA, USA, Cat #M0492L) and the fragments' optimal annealing temperatures (Supplementary Table 1). We purified amplified fragments using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany, Cat #28104). We determined concentration of each synthetic fragment via NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA).

For each fragment designed for methylation, we took 1 µg of synthetic DNA fragment, and methylated using *M.SssI* CpG methyltransferase (Thermo Fisher Scientific, Waltham, MA, USA, Cat #EM0821). We incubated the methylation reaction at 37 °C for 30 min, then 65 °C for 20 min. We purified the methylated product using the MinElute PCR Purification Kit (Qiagen, Hilden, Germany, Cat #28004). To verify methylation, we digested the original PCR amplicon and the methylated PCR amplicon with either HpyCH4IV, HpaII or AfeI restriction enzyme, depending on the cut sites each fragment contained (Supplementary Table 1). We considered methylation successful when, after restriction digest, the PCR amplicon had a single band when run on a 2% agarose electrophoresis gel. Then, using the known relative molecular mass of each synthetic fragment, we determined its molar amount using the Qubit dsDNA HS Assay Kit and a Qubit 2.0 Fluorimeter (Thermo Fisher Scientific, Waltham, MA, USA).

2.3 Assessing technical bias

To assess the performance and any potential biases of synthetic fragments as spike-in controls, we performed cfMeDIP-seq using solely the spike-in control DNA pools as the input. The input pool consisted of 9.99 ng synthetic spike-in DNA, with equimolar amounts of each fragment size within each fragment size pool, and equimolar amounts of each methylation status (Supplementary Table 1).

We performed cfMeDIP-seq as previously described[2] with slight modifications, detailed here. To account for PCR amplification bias, we used xGen Stubby Adapter and unique dual indexing (UDI) primer pairs (IDT, Coralville, IA, USA, Cat #10005924). We performed adapter ligation overnight at 4 °C, adjusting final adapter concentration to 0.09 μmol by dilution.

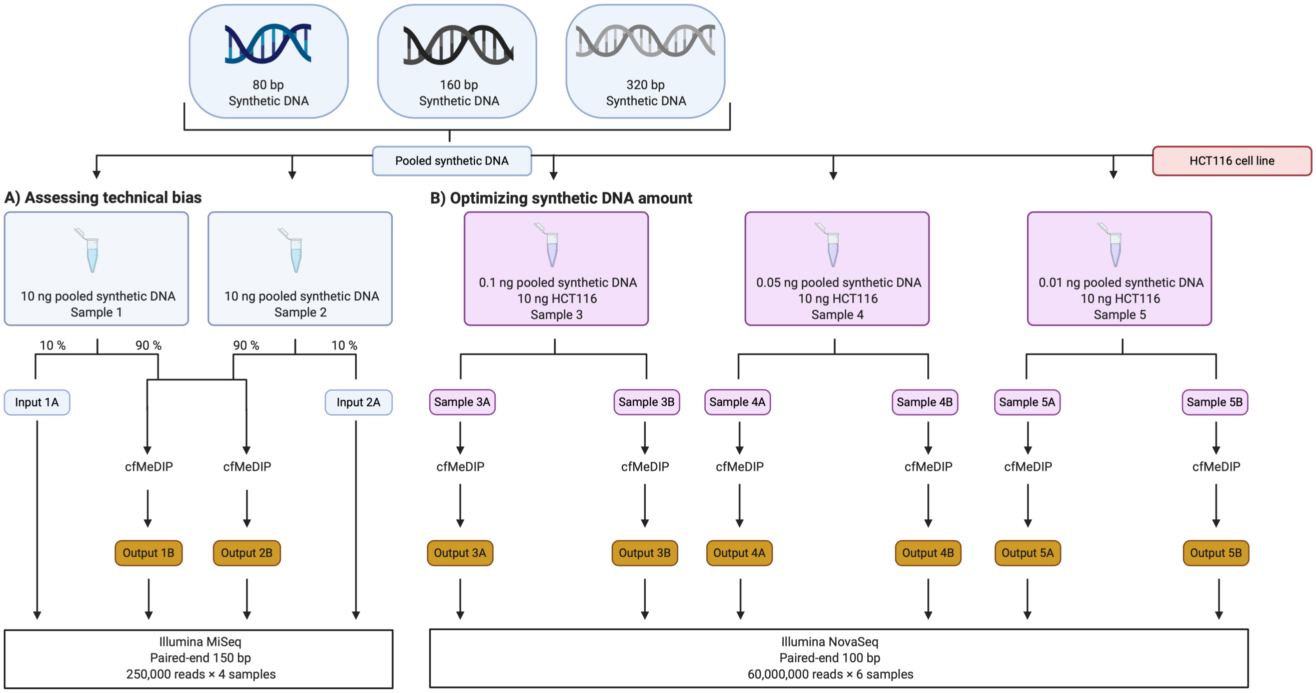

Figure 1: **Experimental design using synthetic spike-in control DNA to (A) assess technical bias and (B) optimize the synthetic DNA amount.** Figure made with BioRender.

For each sample, we saved 10% of the DNA denaturation product as input (Figure 1, samples 1A and 2A). For each sample, we amplified both input and outputs followed by purification and dual size selection using AMPure XP beads (Beckman Coulter, Brea, CA, USA) selecting for fragments between 200 bp–500 bp, including the 120 bp of adapters. These post-ligation fragment sizes reflect an original fragment length between 80 bp–380 bp. Samples underwent sequencing (Princess Margaret Genomics Centre, Toronto, ON, CA) on a MiSeq 2.0 Nano flowcell (Illumina, San Diego, CA, USA), paired-end 2×150 bp, 1 million reads per flowcell (Figure 1).

2.4 Optimizing synthetic DNA amount

We determined the optimal amount of spike-in control DNA needed per experiment by adding varying amounts of spike-in controls to sheared mycoplasma-free HCT116 genomic DNA (American Type Culture Collection, Manassas, VA, USA, RRID: CVCL_0291). Optimizing the amount of spike-in control DNA added to an experiment avoids using a large portion of the sequencing reads on spike-in fragments, saving most reads for the biological sample. We sheared the

HCT116 genomic DNA using an LE220 ultrasonicator (Covaris, Woburn, MA, USA). Using AMPure XP beads (Beckman Coulter, Brea, CA, USA), we size selected to ~150 bp in length to mimic cfDNA input. We created 3 replicate samples of sheared HCT116 cfDNA mimic with masses of synthetic spike-in control DNA of 0.1 ng, 0.05 ng, and 0.01 ng. We performed the cfMeDIP-seq experiment as previously described.[2] Samples underwent sequencing (Princess Margaret Genomics Centre, Toronto, ON, CA) on a NovaSeq 6000 (Illumina, San Diego, CA, USA), paired-end 2×100 bp, 60 million single-end reads per sample (Figure 1).

2.5 Bioinformatic preprocessing

We performed the same bioinformatic preprocessing on all samples from all experiments. We trimmed adapters using fastp version 0.11.5[17] `--umi --umi_loc=per_read --umi_len=5 --adapter_sequence=AATGATACGGCGACCACCGAGATCTACACATATGCGCACACTCTTTCCCTACACGAC --adapter_sequence_r2=CAAGCAGAAGACGGCATACGAGATACGATCAGGTGACTGGAGTTCAGACGTGT`. We removed reads with a Phred score[18] <20, the default of Bowtie2. We aligned reads to the sequences of our designed fragments using Bowtie2[19] version 2.3.5 `bowtie2 --local --minins 80 --maxins 320`, writing unaligned reads to a separate file. We subsequently aligned the previous unaligned reads to our synthetic DNA to the human reference genome (GRCh38/hg38).[14] In every sample, over 98% of reads aligned to either spike-in control sequences or to the human genome. We discarded read pairs when at least one read in the pair did not align or had low quality, defined as Phred score[18] <20. We collapsed fragments with the same UMI files counting them as one fragment. We used samtools[20] version 0.10.2 to convert sequence alignment/map (SAM) files to binary alignment/map (BAM) files.[20]

2.6 Absolute quantification from spike-in control data

We created a Gaussian generalized linear model to predict molar amount from deduplicated spike-in control read counts based on UMI consensus sequence, G+C content, CpG fraction, and fragment length. Using this method, we can absolutely quantify cfDNA. To do this, we used the stats package in R[21] version 3.4.1. Our model estimates molar amount η for each DNA fragment present in the original sample using regression coefficients β learned for each experiment. For each fragment, this model directly includes read count x_{reads}, fragment length x_{len}, G+C content x_{GC}, and CpG fraction $x_{\text{CpG fraction}}$. The final model estimates the molar amount η,

$$\eta = \beta_0 + \beta_{\text{reads}} x_{\text{reads}} + \beta_{\text{len}} x_{\text{len}} + \beta_{\text{GC}} x_{\text{GC}} + \beta_{\text{CpG fraction}} \sqrt[3]{x_{\text{CpG fraction}}}.$$

To reduce the left skew of CpG fraction, we used a cube root transformation.

To calculate the proportion of a given fragment that overlapped with the defined 300 bp windows, we used bedtools version 2.29.2 intersect.[22] We calculated an adjusted molar amount η' to only consider the portion of the window each fragment overlapped. For this calculation, we multiplied the molar amount η by the length of the overlap between the fragment and the genomic window $\ell \in [1\,\text{bp}, 300\,\text{bp}]$, divided by the window size $\ell^* = 300\,\text{bp}$:

$$\eta' = \frac{\ell}{\ell^*}\eta.$$

2.7 Identifying regions to be filtered

To assess multimapping reads that might influence the molar amount estimate, we used Umap[23] multi-read mappability scores. We used k100 mappability scores, representing the largest read lengths available. We annotated each 300 bp window with its minimum mappability score. We assessed the relationship between molar amount and mappability scores.

We calculated standard deviation of molar amount between the two replicate samples for which we spiked 0.01 ng of synthetic DNA into 10 ng of HCT116 genomic DNA. We assessed the relationship between standard deviation of molar amount between replicates, excluding 1 070 387 simple repeat regions,[24] 239 461 regions listed in the Encyclopedia of DNA Elements (ENCODE) Project[25] blacklist,[26] 345 647 regions with mappability score ≤0.5, and 9 587 560 regions with standard deviation ≥0.25. This left 4 551 870 genomic windows in the analyses.

We used HOMER[27] version 4.10.4 to investigate whether specific transcription factor binding motifs associated with molar amount outliers. We compared the outliers to HOMER generated randomized genomic background with window size 300 bp.

2.8 Correlation between picomoles and M-values

We removed simple repeat regions,[24] regions listed in the ENCODE blacklist,[26] regions with Umap k100-multi-read mappability ≤0.5, and regions with standard deviation of molar amount ≥0.25. As described above, we estimated molar amount, using a generalized linear model ($r^2 = 0.93$). We prioritized models that performed better on 160 bp fragments, as we physically size selected for these fragments.

To compare molar amount to another complimentary measure of DNA methylation, we had genomic DNA from the cell line HCT116 profiled (Princess Margaret Genomics Centre, Toronto, ON, CA) using the Infinium MethylationEPIC BeadChip array (Illumina, San Diego, CA, USA). We prepared these samples as technical replicates of the HCT116 genomic DNA that we later spiked with 0.01 ng spike-in control. We normalized and preprocessed array data using sesame[28] version 1.8.2. We annotated CpGs on the array to our 300 bp genomic windows. When >1 CpG probe annotated to a window, we calculated the mean probe M-values across the window.

We assessed correlation between array M-values and picomoles, and between array M-values and read counts. We compared to M-values rather than β values because β values have high heteroscedasticity.[29] As cfMeDIP-seq preferentially enriches for highly methylated regions, we hypothesized that regions for which the array has more CpG probes would correlate to both molar amount and read counts better than regions with less CpG coverage. As such, we assessed the correlation independently at windows containing ≥3, ≥5, ≥7, and ≥10 CpG probes.

2.9 Examining consistency across experimental batches

To experimentally introduce batch effects, we provided a sample of 10 ng of cfDNA obtained from peripheral blood plasma of 5 acute myeloid leukemia (AML) patients containing 0.01 ng of our spike-in controls to 3 independent labs. We obtained the deidentified patient samples, previously included in Shen *et al.*,[1] from the Leukemia Tissue Bank, Princess Margaret Cancer Centre, University Health Network with informed consent following approval by the University Health Network Research Ethics Board (01-0573). Blood was collected at the time of diagnosis in EDTA tubes. Samples were spun and plasma frozen in Eppendorf tubes at –80 °C until use. As a blood cancer, leukemia generates a high amount of plasma cfDNA, allowing us to have sufficient cfDNA to divide into three technical replicates of 10 ng each.

Each lab performed the cfMeDIP-seq method described by Shen et al,[2] with a variety of procedural modifications, detailed here. The changes emulated batch effects that would be commonly seen in publicly available data from different labs for different studies. Labs 1 and 3 used the same IDT xGen Duplex Seq adapters with 3 bp UMI, as described above. Lab 2 used custom IDT adapters with 2 bp degenerate UMI, as previously described.[30] For ligation of adapters, Labs 1 and 2 incubated at 4 °C for 16 h, while Lab 3 incubated at 20 °C for 2 h. Labs 1 and 3 used Antibody 1 (Diagenode, Denville, NJ, USA, Cat #C15200081-100, Lot #RD004, RRID: AB_2572207), while Lab 2 used Antibody 2 (Diagenode, Denville, NJ, USA, Cat #C15200081-100, Lot #RD001, RRID: AB_2572207). Lab 1 and 2 used 15% methylated lambda filler DNA and 85% unmethylated lambda filler DNA. Lab 3 used 100% unmethylated lambda filler DNA.

For amplifying the final library, the batches had different numbers of PCR cycles. Lab 1 ran 13–15 cycles, optimized per sample, Lab 2 ran 13 cycles, and Lab 3 ran 11 cycles. Lab 1 and 2 sequenced DNA with a NovaSeq 6000 (Illumina, San Diego, CA, USA), paired-end 2×100 bp. Lab 3 sequenced DNA on a NextSeq 550 (Illumina, San Diego, CA, USA), paired-end 2×75 bp. Lab 1 obtained 60 million reads per sample, Lab 2 obtained 100 million reads per sample, and Lab 3 obtained 85 million reads per sample.

We assessed whether spike-in controls mitigated batch effects on samples for which we calculated molar amount.

To do this, we performed principal component analysis (PCA) on four different quantification methods:

1. read counts only, without the use of spike-in controls
2. read counts preprocessed using QSEA version 1.16.0,[31] the current standard processing pipeline for the methylated DNA immunoprecipitation-sequencing (MeDIP-seq) data
3. molar amount, without filtering
4. molar amount, with filtering

To investigate whether known variables associated with any of the principal components, we performed two-way analysis of variance (ANOVA) between each principal component and each categorical variable. Categorical variables included: batch, sequencing machine, adapters, samples, and sex (inferred by Y chromosome signal). We converted the resulting F-statistic to an effect size, Cohen's d,[32] using the compute.es package[33] version 0.2.5 and R version 3.4.1. We adjusted p-values for multiple test correction using the Holm-Bonferroni method.[34]

3 Results

3.1 Spike-in controls confirm cfMeDIP's high efficiency and specificity

We performed cfMeDIP-seq directly on the synthetic spike-in control fragments. For each sample, we saved 10% of the mass before performing cfMeDIP-seq. This acts as an input control (Figure 1). In input samples, we observed 51% of the input fragments methylated and 49% unmethylated. After cfMeDIP, 97% of the output fragments were methylated (Figure 2A,B). The enrichment for methylated sequences further supports the validity and high efficiency of cfMeDIP-seq.

After cfMeDIP, signal from methylated fragments for both the synthetic spike-in control alone (97% of spike-in control fragments) and 10 ng HCT116 DNA with 0.01 ng spike-in (99.99% of spike-in control fragments) showed an enrichment of 160 bp fragments. We expected this enrichment due to our size selection step for insert size fragments between 80 bp–380 bp. We also observed enrichment of fragments with higher G+C content and higher CpG fraction (Figure 2B,C).

Signal from unmethylated fragments for both the synthetic spike-in control alone (3% of spike-in control fragments) and 10 ng HCT116 DNA with 0.01 ng spike-in (0.01% of spike-in control fragments) had no association with fragment lengths, G+C content, or CpG fraction (Figure 2B,C). This suggests the low amount of unmethylated fragment signal arose from random non-specific binding and was not confounded systematically with fragment length, G+C content, or CpG fraction.

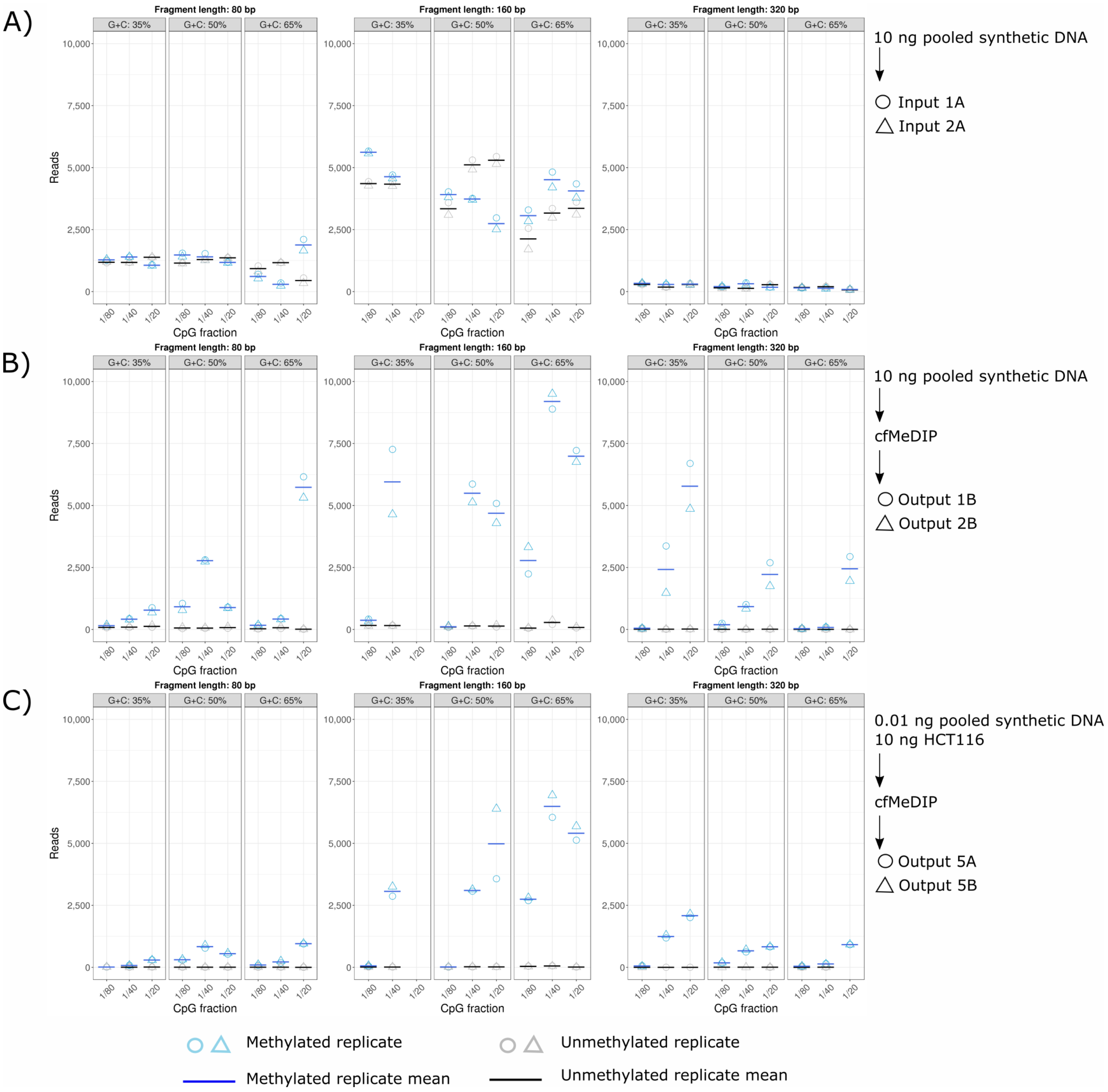

Figure 2: **Assessing biases in fragment length, G+C content, and CpG fraction in (A) input spike-in control DNA without cfMeDIP-seq, (B) output spike-in control DNA, after cfMeDIP-seq and (C) 0.01 ng spike-in control DNA added to HCT116 replicates.** Blue: methylated fragments; grey: unmethylated fragments. Circle: sample 1; triangle: sample 2. Solid line: mean of the two samples.

3.2 Low-input spike-in control accounts for technical variance

To determine the optimal amount of spike-in control DNA to be added to each sample, we assessed the the proportion of reads used towards our spike-in controls. We compared spike-in control reads to the total number of reads used for our biological sample, HCT116 genomic DNA. We optimized the amount of spike-in controls to be used in subsequent experiments. This allowed us to maximize reads to our biological sample of interest while obtaining sufficient reads from the spike-in controls to correct for technical bias. Adding in 0.01 ng of our synthetic spike-in control DNA before cfMeDIP-seq used <1% of the reads for the spike-in controls. We retained >650 000 reads of spike-in control sequence for analysis while leaving the rest of the reads for our biological sample. Therefore, we decided to use 0.01 ng of spike-in control fragments in subsequent experiments.

The 0.01 ng of spike-in control DNA added to 10 ng of HCT116 genomic DNA revealed ≥99.99% specificity of methylated DNA with ≤0.01% total non-specific binding to non-methylated fragments (Figure 2C). We calculated methylation specificity by dividing the total number of methylated fragments by the total number of spike-in control fragments. The cfMeDIP process also enriched for the 160 bp fragments we physically size selected for and for higher G+C content. Fragments that have CpG present at only 1/80 bp in the experiment with 10 ng of HCT116 with 0.01 ng of spike-in control DNA were represented by ≤1% of reads (Figure 2C). The same patterns persisted when spiking in 0.05 ng and 0.1 ng spike-in control DNA into 10 ng of HCT116 cell line.

3.3 Removing problematic regions eliminates some technical artifacts

From our normalized data, we removed regions containing simple repeats, the ENCODE blacklist regions, and regions with mappability scores ≤0.5. We noticed that many regions with high molar amount also had high standard deviation of molar amount between replicate samples. Thus, we removed regions where standard deviation of molar amount between replicates was ≥0.25.

After removing the high standard deviation regions, we observed no relationship between molar amount and standard deviation of molar amount, and no relationship between molar amount and mappability. This suggests that removing regions that have high standard deviation of molar amount between replicates can reduce technical artifacts. Regions with high standard deviation of molar amount between replicates perform inconsistently, leading to inaccurate quantification of DNA methylation.

Eleven genomic windows associated with high predicted molar amount (Figure 3). These regions all had molar amount ≥2 pmol (Table 1). All of these 11 windows match repetitive elements, predominantly short interspersed nuclear elements (SINEs), mostly from the Alu family. Most of the Alu elements belonged to the older S and J subfamilies[36] (Ta-

Chrom[a]	Start[b]	End[b]	Amount	Repeat element[c]	Repeat family[c]	Repeat name[c]
13	95 176 201	95 176 500	78.2 pmol	SINE	Alu	AluJo
2	120 383 401	120 383 700	55.0 pmol	SINE	MIR	MIR_Amn
12	95 476 201	95 476 500	11.7 pmol	SINE	Alu	AluSx
22	20 900 401	20 900 700	6.3 pmol	SINE	Alu	AluJb, AluY
17	1 025 101	1 025 400	5.3 pmol	SINE	Alu	AluYe5, AluSx1
X	44 613 001	44 613 300	4.2 pmol	SINE	Alu	AluSp, AluJr
1	44 582 701	44 583 000	3.8 pmol	SINE	Alu	AluSx1
8	139 704 301	139 704 600	3.7 pmol	Low complexity	—	G-rich
2	224 230 501	224 230 800	2.5 pmol	LTR	ERV1	HERVH-int
17	3 521 101	3 521 400	2.5 pmol	LINE, DNA transposon	L2, hAT-Blackjack	L2b, MER63D
16	11 578 801	11 579 100	2.4 pmol	LINE, SINE	L1, Alu	L1MD3, AluSp

Table 1: **The 11 genomic windows of length 300 bp with predicted molar amount ≥2 pmol.** Sorted by decreasing molar amount.

[a] Chromosome.

[b] GRCh38/hg38, genomic position 1-start, fully closed.

[c] All RepeatMasker[35] version 3.0 track repeat elements, families, and names that overlap our 300 bp genomic windows.

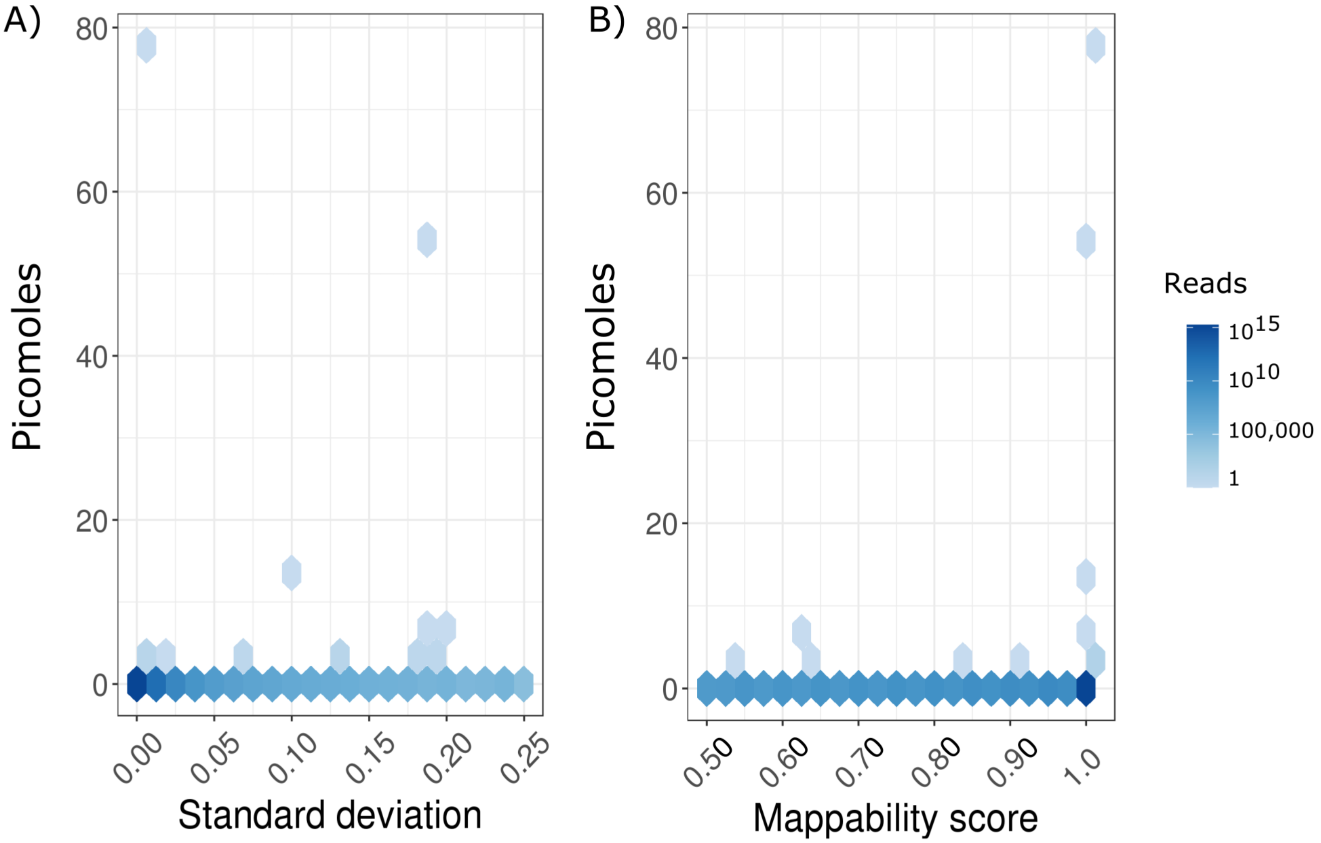

Figure 3: **Two-dimensional histogram of the number of reads found in 300 bp windows, as binned by molar amount and either (A) standard deviation of molar amount or (B) Umap k100 multi-read mappability.** Only includes windows that do not overlap with University of California, Santa Cruz (UCSC) simple repeats and the ENCODE blacklist, and regions with Umap k100 multi-read mappability scores ≤0.5.

ble 1). We searched HOMER[27] for transcription factor binding motifs associated with molar amount, but found no significant motifs.

3.4 Absolute quantification correlates with M-values

We compared molar amount to M-values from the EPIC array. We used our generalized linear model to calculate molar amount of cfMeDIP-seq methylated DNA fragments for each 300 bp genomic window. Molar amount significantly correlated with array M-values over 300 bp in our HCT116 genomic DNA samples ($r \geq 0.63$; Figure 4A,C,E,G). Correlation increased when we restricted analyses to 300 bp windows with ≥5 CpG probes on the array ($r = 0.79$; Figure 4C). This reflected cfMeDIP-seq's preference for methylated, CpG-dense reads.

We compared the current standard of read counts to M-values (Figure 4B,D,F,H). Molar amount correlated with M-values similarly to read counts, but provides the advantage of absolute quantification.

3.5 Spike-in controls significantly mitigate batch effects

To determine whether our spike-in controls mitigate batch effects, we provided aliquots of cfDNA samples from 5 AML patients to three different labs. Each lab performed cfMeDIP-seq on each of the 5 samples with slight variations.

We performed PCA to assess if any batch variables drive any of the top principal components. Principal component 1 explains 78% of the variance and associated with processing batch for raw read count data without spike-in controls (Figure 5A). QSEA normalization only reduced the variance explained by principal component 1 to 75%. Even with QSEA normalization, principal component 1 associated with the batch variable unmethylated lambda filler DNA (Figure 5B). The correlation of principal component 1 to unmethylated filler DNA only appeared after QSEA normalization, suggesting that QSEA may have introduced stochastic variance into the data. QSEA normalization principal component 1 still associated with batch, although not statistically significantly after multiple test correction.

Using spike-in controls greatly reduced batch effects, even with significantly different protocols between batches. Quantifying the data by molar amount generated with all genomic regions shifted the association of batch processing variable to principal component 2, explaining ≤5% of variance (Figure 5C). Our suggested filtering, including removing simple repeats, the ENCODE blacklist regions, and regions with low mappability, further shifted the association of batch processing to principal component 5, explaining only 1% of variance (Figure 5D).

We further investigated principal component 5 using spike-in controls and removing simple repeats, the ENCODE blacklist regions and regions with low mappability. Examining the top 10% of windows ($N = 258\,254$) driving the explained variance, 72% matched to RepeatMasker[35] repetitive elements. Of the mapped repetitive elements, 72% mapped

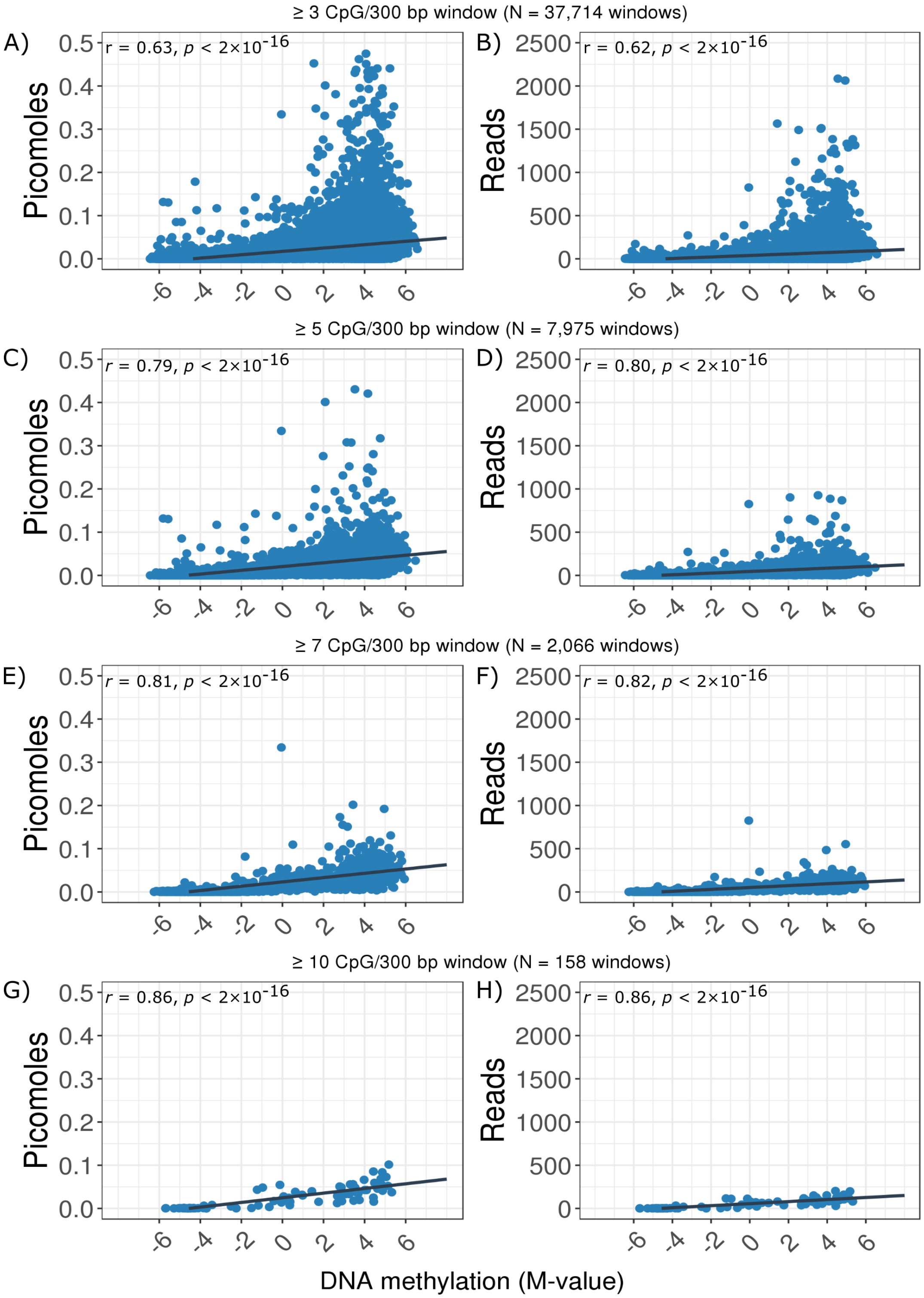

Figure 4: **Correlation of two measurements of fragment methylation by cfMeDIP and EPIC array M-value for 300 bp genomic windows. (A,C,E,G) Molar amount calculated from HCT116 samples correlated to EPIC array M-values. (B,D,F,H) Read counts calculated from the same samples, ignoring the spike-in controls. (A,B) 37 714 windows with ≥3 CpG probes represented on the EPIC array. (C,D) 7975 windows with ≥5 CpG probes represented on the EPIC array. (E,F) 2066 windows with ≥7 CpG probes represented on the EPIC array. (G,H) 158 windows with ≥10 CpG probes represented on the EPIC array.**

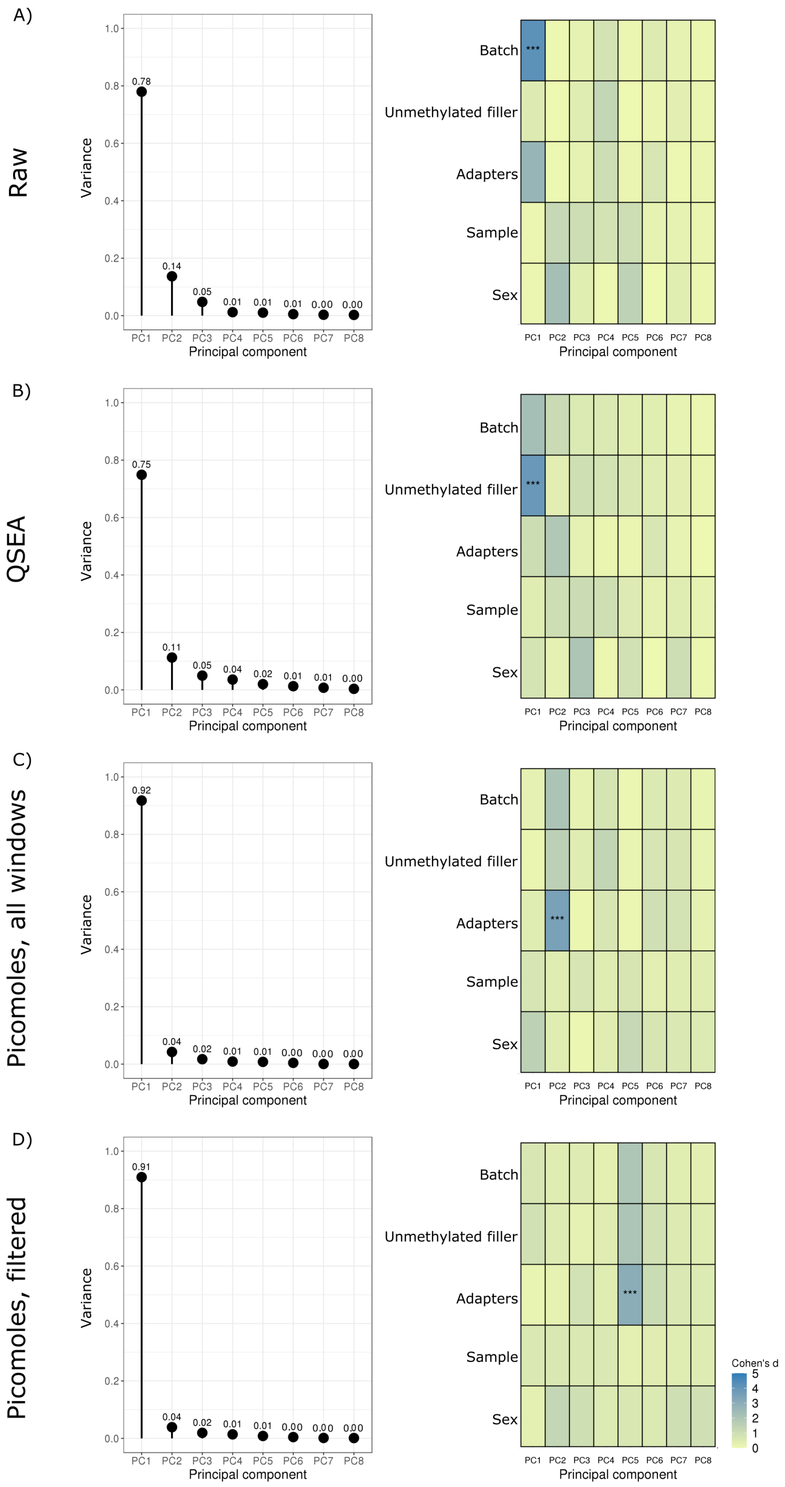

Figure 5: **Principal component analyses of cfMeDIP results normalized through 4 different strategies, and associations with experimental variables.** *(Left)* Proportion of the variance explained by each principal component. *(Right)* Association between known variables, both technical and clinical, and principal component. Cohen's d is an effect size of standardized means between variable. *** $p < 0.001$. **(A) Raw read counts without any normalization. (B) Read counts normalized using QSEA. (C) Data normalized using spike-in controls. (D) Data normalized using spike-in controls and removing regions in UCSC simple repeats, in the ENCODE blacklist, and with Umap k100 multi-read mappability scores ≤0.5.**

to Alu elements. Batch effects associated with repetitive elements suggest inconsistent or inaccurate measures of DNA methylation at these regions.

4 Discussion

The data above establish the validity of using our synthetic spike-in control DNA to absolutely quantify cfDNA in cfMeDIP-seq experiments. We showed that technical bias exists in the cfMeDIP-seq data, and that the use of our spike-in controls helps mitigate these biases. In cfMeDIP-seq data not using the spike-in controls, the batch effects may be stark as Lab 3 used unmethylated lambda filler DNA. The spike-in controls successfully mitigated the effect of this change. One can use the signal from unmethylated spike-in control reads to calculate methylation specificity for each sample. To reduce technical artifacts, one must remove problematic genomic regions prior to analysis.

Despite removing problematic genomic regions, some remaining regions had much higher predicted molar amount than all other genomic regions. The high molar amount genomic regions consisted mostly of repetitive elements, predominantly SINEs (Table 1). While these regions had high CpG density, our model already adjusted for CpG fraction. This made CpG density an unlikely driver of high molar amount.

We may see over-representation of high molar amount in some repetitive regions due to the characteristic hypermethylation of these regions.[36] Had we not removed many repetitive regions when removing regions listed in the ENCODE blacklist and regions with low mappability, we may have observed more genomic regions with predicted high molar amount. Regions with high molar amount that map to repetitive elements contain many extra copies not present in the reference genome. These regions would appear uniquely mappable, and thus not removed by our previous filtering steps.

HCT116 likely has problematic genomic regions not found in the ENCODE blacklist. This arises from the relative dearth of ENCODE data available for the blacklist generation process, compared to cell types in the ENCODE Tier 1 and Tier 2 categories.

Depending on the experimental question, some may choose to go beyond our filtering recommendations and remove all repetitive elements, such as all long interspersed nuclear elements (LINEs) and all SINEs. Given that, after filtering, only 11 genomic windows had molar amount ≥2 pmol, removing all repetitive elements would not affect results drastically.

Both molar amount in picomoles and read counts correlated to M-values. The strength of this correlation varied with the number of CpGs represented on the EPIC array for a genomic window. We expect the increase in correlation with more CpGs as cfMeDIP-seq preferentially enriches for hypermethylated regions, and picks up fewer fragments in

CpG-sparse regions. Additionally, if the array has fewer probes representing a 300 bp genomic region than that region has CpGs, it may poorly represent DNA methylation for the entire 300 bp.

To facilitate the use of our spike-in controls, we have created an R package, `spiky`, to help process data generated from cfMeDIP-seq experiments that include the spike-in controls. This package trains the Gaussian generalized linear model and predicts molar amount in picomoles on user data. The `spiky` package is available on GitHub (`https://github.com/trichelab/spiky`) and will soon be available on Bioconductor.[37]

Incorporating these spike-in controls in future cfMeDIP-seq experiments will adjust for technical biases and mitigate batch effects, improving cfMeDIP-seq data overall.

Data availability

We deposited processed data and raw non-human data in the Gene Expression Omnibus[38] (GEO accession: GSE166259). We deposited raw data for the AML samples in the European Genome-phenome Archive[39] (EGA).

Code availability

The `spiky` package is available on GitHub (`https://github.com/trichelab/spiky`) under the GNU General Public License version 2 license. Other scripts are available on GitHub (`https://github.com/hoffmangroup/2020spikein`) and deposited on Zenodo (`https://doi.org/10.5281/zenodo.4533340`).

Author contributions

Conceptualization, S.L.W., S.Y.S., D.D.De C., and M.M.H.; Data curation, S.L.W. and S.Y.S.; Formal analysis, S.L.W.; Funding acquisition, S.L.W., S.V.B., D.D.De C., and M.M.H.; Investigation, S.Y.S. and J.B.; Methodology, S.L.W., S.Y.S., T.T., D.D.De C., and M.M.H.; Project administration, S.L.W. and M.M.H.; Resources, M.M.H.; Software, S.L.W., L.H., and T.T.; Supervision, S.V.B., T.T., D.D.De C., and M.M.H.; Validation, S.L.W. and S.Y.S.; Visualization, S.L.W.; Writing — original draft, S.L.W.; Writing — review & editing, S.L.W., S.Y.S, L.H., J.B., T.T., S.V.B., D.D.De C., and M.M.H.

Acknowledgments

We thank Dax Torti (0000-0002-9000-5803) (Ontario Institute for Cancer Research) for performing batch analysis experiments, Mark D. Minden (0000-0002-9089-8816) and Andrea Arruda (0000-0003-2516-8005) (Leukemia Tissue

Bank, Princess Margaret Cancer Centre, University Health Network) for providing the AML samples and for comments on the manuscript, and Zhibin Lu (0000-0001-6281-1413) (Bioinformatics and High Performance Computing Core, University Health Network) for technical assistance. This work was supported by the Canadian Institutes of Health Research (389866 to M.M.H., MFE-171256 to S.L.W.), the McLaughlin Centre (MC-2019-07 to M.M.H.), the Molly Towell Perinatal Research Foundation, and the Princess Margaret Cancer Foundation. T.T. is supported by startup funds from the Van Andel Institute, the Michigan Economic Development Corporation, the Michelle Lunn Hope Foundation, and the Folz Family Fund for Cancer Research.

Competing interests

S.L.W., S.Y.S., T.T., D.D.De C., and M.M.H. are inventors on a patent application related to the synthetic spike-in controls. S.Y.S., S.V.B., and D.D.De C. are inventors on other patent applications related to this work. S.V.B. and D.D.De C. are co-founders of and provide consulting for DNAMx, Inc. S.V.B. and D.D.De C. have received research funding from Nektar Therapeutics.

References

1. Shen, S. Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M. H., Chadwick, D., Zuzarte, P. C., Borgida, A., Wang, T. T., Li, T., *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563,** 579–583 (2018).
2. Shen, S. Y., Burgener, J. M., Bratman, S. V. & De Carvalho, D. D. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nature Protocols* **14,** 2749–2780 (2019).
3. Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R. & Oliver, B. Synthetic spike-in standards for RNA-seq experiments. *Genome Research* **21,** 1543–1551 (2011).
4. Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W. & Tyler, J. K. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Molecular and Cellular Biology* **36,** 662–667 (2016).
5. Orlando, D. A., Chen, M. W., Brown, V. E., Solanki, S., Choi, Y. J., Olson, E. R., Fritz, C. C., Bradner, J. E. & Guenther, M. G. Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Reports* **9,** 1163–1170 (2014).
6. Deveson, I. W., Chen, W. Y., Wong, T., Hardwick, S. A., Andersen, S. B., Nielsen, L. K., Mattick, J. S. & Mercer, T. R. Representing genetic variation with synthetic DNA standards. *Nature Methods* **13,** 784–791 (2016).

7. Blackburn, J., Wong, T., Madala, B. S., Barker, C., Hardwick, S. A., Reis, A. L., Deveson, I. W. & Mercer, T. R. Use of synthetic DNA spike-in controls (sequins) for human genome sequencing. *Nature Protocols* **14,** 2119–2151 (2019).

8. Plongthongkum, N., Diep, D. H. & Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews Genetics* **15,** 647–661 (2014).

9. Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Böckler, B. & Song, C.-X. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature Biotechnology* **37,** 424–429 (2019).

10. Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R. & Reik, W. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biology* **19,** 33 (2018).

11. Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., Mair, R., Goranova, T., Marass, F., Heider, K., *et al.* Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine* **10,** e4921 (2018).

12. Ponty, Y., Termier, M. & Denise, A. GenRGenS: software for generating random genomic sequences and structures. *Bioinformatics* **22,** 1534–1535 (2006).

13. Madden, T. The BLAST sequence analysis tool (2013).

14. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., *et al.* Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27,** 849–864 (2017).

15. Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., McQuisten, K. A., Almabrazi, H. G., Pedersen, K. F., Lin, Y., Garretson, J., McEntaggart, N. O., *et al.* IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Research* **36,** e163 (2008).

16. Xu, Z. Z. & Mathews, D. H. Secondary structure prediction of single sequences using RNAstructure. *RNA Structure Determination,* 15–34 (2016).

17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34,** i884–i890 (2018).

18. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research* **8,** 186–194 (1998).

19. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9** (2012).

20. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

21. R Core Team. *R: A Language and Environment for Statistical Computing,* R Foundation for Statistical Computing (Vienna, Austria, 2013).

22. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

23. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research* **46,** e120 (2018).

24. Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. & Kent, W. J. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32,** D493 (2004).

25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57 (2012).

26. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Scientific Reports* **9,** 9354 (2019).

27. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime *cis* regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38,** 576–589 (2010).

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

29. Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L. & Lin, S. M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587 (2010).

30. Wang, T. T., Abelson, S., Zou, J., Li, T., Zhao, Z., Dick, J. E., Shlush, L. I., Pugh, T. J. & Bratman, S. V. High efficiency error suppression for accurate detection of low-frequency variants. *Nucleic Acids Research* **47,** e87 (2019).

31. Lienhard, M., Grasse, S., Rolff, J., Frese, S., Schirmer, U., Becker, M., Börno, S., Timmermann, B., Chavez, L., Sültmann, H., *et al.* QSEA—modelling of genome-wide DNA methylation from sequencing enrichment experiments. *Nucleic Acids Research* **45,** e44 (2017).

32. Rice, M. E. & Harris, G. T. Comparing effect sizes in follow-up studies: ROC, Cohen's *d*, and *r*. *Law and Human Behavior* **29,** 615–620 (2005).

33. Del Re, A. C. *compute.es: Compute Effect Sizes* (2013).

34. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70 (1979).

35. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.* 2015.

36. Deininger, P. Alu elements: know the SINEs. *Genome Biology* **12,** 236 (2011).

37. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12,** 115–121 (2015).

38. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30,** 207–210 (2002).

39. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics* **47,** 692–695 (2015).