Comparison and imputation-aided integration of five commercial platforms for targeted DNA methylome analysis

Miljana Tanić1,2, Ismail Moghul1, Simon Rodney1, Pawan Dhami1,3, Heli Vaikkinen3,4, John Ambrose5, James Barrett1, Andrew Feber1,6,7 and Stephan Beck1,8

Targeted bisulfite sequencing (TBS) has become the method of choice for the cost-effective, targeted analysis of the human methylome at base-pair resolution. In this study, we benchmarked five commercially available TBS platforms—three hybridization capture-based (Agilent, Roche and Illumina) and two reduced-representation-based (Diagenode and NuGen)—across 11 samples. Two samples were also compared with whole-genome DNA methylation sequencing with the Illumina and Oxford Nanopore platforms. We assessed workflow complexity, on/off-target performance, coverage, accuracy and reproducibility. Although all platforms produced robust and reproducible data, major differences in the number and identity of the CpG sites covered make it difficult to compare datasets generated on different platforms. To overcome this limitation, we applied imputation and show that it improves interoperability from an average of 10.35% (0.8 million) to 97% (7.6 million) common CpG sites. Our study provides guidance on which TBS platform to use for different methylome features and offers an imputation-based harmonization solution that allows comparative, integrative analysis.

DNA methylation is an indispensable epigenetic mark for many biological processes, such as development, differentiation and maintenance of cell-type-specific states. Interrogating the changes in DNA methylation patterns is essential to better understand the biology of normal and pathologic states and to identify clinically relevant biomarkers. In the past decade, methods for methylome analysis have moved away from semi-quantitative methods with coarse resolution (MeDIP and MRE-seq) toward methods with the single-base resolution based on bisulfite conversion, such as microarrays and next-generation sequencing (NGS). NGS offers several advantages, including single-molecule analysis and read phasing to sample heterogeneity and epiallele composition. Whole-genome bisulfite sequencing (WGBS) remains the gold standard method for studying DNA methylation at a single base pair (bp) resolution, although it is associated with high costs and requires considerable computational resources. Targeted bisulfite sequencing (TBS) directs the sequencing to more informative parts of the genome, reducing the costs. The analysis of DNA methylation by TBS can be achieved either through target-specific enrichment of regions containing CpG sites of interest using probe hybridization capture (HC) methods or through non-specific enrichment of CpG-dense regions by reduced-representation bisulfite sequencing (RRBS) mediated by a restriction enzyme recognition site containing a CG motif.

Currently, five commercial manufacturers are producing off-the-shelf kits, offering standardized reagents and conditions for genome-wide TBS using HC (Agilent SureSelect Methylation-Seq, Roche NimbleGen SeqCap EpiGiant and Illumina TrueSeq Methylation Capture EPIC) or RRBS (Diagenode Premium RRBS and NuGen Ovation RRBS Methyl-Seq). The five platforms employ different experimental strategies to generate sequencing libraries and differ in the scope of regions they cover. The diversity of these platforms’ characteristics, along with the absence of a thorough comparison of their output and performance, warrants a comprehensive benchmarking to provide guidance for users to select the most appropriate platform based on each one’s strengths and limitations.

Here we systematically compare performance of these five platforms in terms of sequencing output, target capture efficiency, genomic features coverage, CpG coverage similarity, intra-platform reproducibility, between-platform concordance and differential methylation, across a set of 11 samples. We benchmark each of the five TBS platforms to gold standard data generated using Illumina WGBS and Nanopore sequencing. Finally, we evaluate interoperability of the TBS platforms and provide guidance for their harmonization.

Results

DNA methylation data generation. To ensure reproducibility and data availability under open-access agreement, we selected for the study a set of 11 well-characterized and/or commercially sourced samples. These included common reference standards: Coriell-NA12878 and HeLa cell lines processed in duplicate; a human reference genomic DNA (gDNA) sample isolated from pooled healthy peripheral blood (Ref.gDNA), representing a heterogeneous mix of cell types, at two different DNA inputs (recommended by the manufacturer and 500ng) in duplicate; four DNA methylation standards generated from commercial (ZYMO) fully methylated and unmethylated control samples; a pair of genetically
and phenotypically divergent bladder cancer cell lines (T24 versus 253J); and a pair of isogenic bladder cancer cell lines with different sensitivity to cisplatin treatment (RT112 versus RT112-CP) selected to reflect two extremes for differential methylation calling. For each TBS platform, a total of 16 libraries were generated (Fig. 1a) using the same source DNA from 11 biological replicates, generating a total of 80 TBS libraries (Supplementary Table 1). The length and complexity of the library prep protocols vary considerably among the methods (Fig. 1b). Libraries were sequenced at 100 bp PE, and data were processed using the same computational pipeline. Additionally, new WGBS data were generated for Ref.gDNA and Oxford Nanopore for Coriell-NA12878.

**Differences in platform design.** The five platforms differ in the methods used for DNA fragmentation, type of hybridization probes, total size of the epigenome covered, number of targeted CpGs, strand specificity, required DNA input, protocol complexity, total hands-on time, time-to-result and, finally, price per sample (Table 1).

Although HC platforms target specific regions of the genome by design, RRBS non-specifically enriches for CpG-dense sequences. To compare the overlap between the sizes of the regions covered by each platform, for HC platforms we used manufacturers’ design files, and for RRBS platforms we performed an in silico enzymatic digestion with size selection mimicking experimental and sequencing conditions to determine the theoretical coverage. All platforms targeted in common only 9,260,583 bp, and 18,204,312 bp were shared by all HC platforms, whereas Illumina and in silico RRBS targeted the most platform-specific regions: 56,996,695 bp and 67,747,965 bp, respectively (Fig. 1c). The size of target regions determines the amount of sequencing required but does not necessarily correlate with the number of enclosed CpG sites. The number of CpGs targeted by each platform ranged from 2.8 million to 4 million CpGs. A total of 707,686 CpGs were commonly targeted by all platforms (Fig. 1d).

**Sequencing output variability.** Sequencing cost-efficiency depends on several factors, including read quality, mapping rates, duplication rates and, in case of hybridization capture methods, target capture efficiency. To simulate real-life biological experimental setup and to enable an unbiased comparison between platforms, equimolar amounts of each sample library were pooled together and sequenced to achieve an average of 30 million uniquely mapped de-duplicated reads (UMDRs). We compared the quality of raw and pre-processed data (Extended Data Fig. 1 and Supplementary Fig. 1) and the technical variability of sequencing output (Fig. 2a and Supplementary Table 2). To achieve the targeted number of UMDRs, between ~55 million reads (for NuGen) and over ~80 million reads (for Illumina) had to be sequenced. Due to reduced sequence diversity of bisulfite-converted DNA, alignment rates are generally lower than for genome sequencing. Mapping rates for HC-based methods were similar between platforms (~79%) and higher than for RRBS-based platforms (~64%). Bisulfite conversion (BC) is a harsh chemical process that degrades DNA, reducing the library complexity and, if combined with a high number of polymerase chain reaction (PCR) cycles, may lead to a large number of duplicate reads skewing the methylation readout49. De-duplication can be achieved by relying on fragment genomic coordinates40 (Agilent, Roche and Illumina) or using unique molecular identifiers (UMIs)42 (Nugen) but not for RRBS without UMIs (Diagenode). Duplication rates were the highest for Illumina’s platform (~52%) and the lowest for NuGen’s (~7%). Overall, the highest ratio of UMDRs to passing filter (PF) reads was observed for NuGen’s platform and the lowest for Illumina’s.

**Target capture efficiency.** Sequencing cost-effectiveness for HC methods also depends on the percentage of sequencing directed to regions of interest versus the amount of off-target sequencing. For this analysis, we omitted RRBS-based platforms because their enrichment is not target specific. The performance of the wet-lab protocol for the three HC platforms’ on-target capture efficiency was evaluated based on coverage of all targeted bases according to manufacturers’ design files (Supplementary Table 3). Overall, on-target capture efficiency measured as the percentage of uniquely aligned PF bases that mapped on or near bait was highest for Illumina’s platform (~90.6%), followed by Agilent’s (~78.2%) and Roche’s (~61.5%), with the two latter platforms marked by high inter-sample variations (Fig. 2b). Fold enrichment by which the baited region has been amplified above the genomic background was similar between platforms (21.3–24.3) (Fig. 2c).

Target coverage was assessed as a proxy for how well the data are likely to perform in downstream applications. Mean target coverage was similar for all platforms, from 35.4× (Roche) to 37.4× (Illumina) (Fig. 2d). We observed higher uniformity of coverage for Roche’s and Illumina’s platforms (~2.3 and ~2.4, respectively) compared to Agilent’s (~3.3) (Fig. 2e). The uniformity of coverage translates into the fraction of targeted regions being covered at a specific depth (Extended Data Fig. 2). Intra-platform variability was pronounced for all three HC platforms, including technical replicates, with the average percentage of targets covered ≥10x ranging from ~80.9% for Agilent to ~88.1% for Illumina (Fig. 2f).

**CpG coverage.** An important consideration when choosing a platform for methylome analysis is the number of CpG sites it interrogates. The number of CpGs covered by RRBS is a function of the depth of sequencing and sequencing conditions (read length and single-end or paired-end mode). Although each MspI-generated fragment has a CpG site at its ends, the number of CpGs within each fragment is not uniform due to the random selection of recognition sites by the enzyme. In contrast, for HC methods, the number of CpGs covered is determined by design. However, we observed a much higher number of experimentally covered CpGs at 1× than expected for all three HC platforms, consistent with high levels of off-target sequencing (Fig. 3a). The average number of CpGs sites at 5× drops to ~4 million CpGs for all platforms but remains higher for RRBS platforms at 10× and 30× compared to HC methods. Both RRBS platforms showed similar CpG coverage, ranging from over 5 million CpGs at 1× to ~2 million CpGs at 30×. Mean CpG depth of coverage was higher for RRBS methods (Diagenode ~33× and NuGen ~31×) compared to HC methods (Agilent ~13×, Roche ~10× and Illumina ~13×), reflecting a more uniform coverage for the former (Supplementary Table 2).

Next, we determined the degree of CpG coverage similarity between and within platforms, in terms of the number of overlapping CpGs covered ≥10x. For each sample platform, we calculated a pairwise Jaccard similarity index defined as the number of commonly interrogated CpG sites (intersection) over the total number of CpGs (union). Unsupervised hierarchical clustering indicated that, by and large, RRBS-based and HC-capture-based platforms cover different CpG sites (Fig. 3b). All samples clustered according to the platform of origin, with Diagenode showing the highest degree of intra-platform similarity. As expected, NuGen and Diagenode RRBS platforms covered mostly the same CpG sites, whereas, among HC methods, Agilent and Roche platforms had a higher degree of similarity compared to Illumina.

**Theoretical and empirical coverage of genomic features.** The CpG distribution is not uniform throughout the genome; thus, an important consideration for choosing a platform is the coverage of genomic features of interest. HC platforms target CpG islands (CGIs), shores and shelves, regulatory regions and known differentially methylated regions (DMRs) to a different extent by design. We compared the size of targeted genomic features and the number of...
Fig. 1 | Technology and design comparison of TBS platforms. a, Schematic outline of the study design for TBS benchmarking. b, Schematic overview of the library preparation protocol steps for each of the platforms. c, UpSet plot showing the size of the intersection of regions targeted by design for Agilent, Roche and Illumina hybridization capture platforms and in silico predictions for RRBS. d, Venn diagram showing the overlap of CpG sites targeted by design for Agilent, Roche and Illumina hybridization capture platforms and in silico predictions for RRBS.
CpGs targeted in each platform (Supplementary Fig. 2). Illumina’s platform targets ~20 megabases (Mb) more than other platforms, including the highest portion of FANTOM5 (ref. 1)–defined enhancers (>40%) and insulator regions. Agilent’s platform targets the largest region and the highest number of CpGs within CGIs, promoters and exons. Interestingly, although Illumina’s platform targets the largest portion of inter-CGI, intron, and heterochromatin regions, RRBS encompasses more CpGs in those regions.

To determine how well each platform experimentally covers these regions, we counted the number of CpGs covered ≥10× per platform in each genomic/regulatory feature (Fig. 3c). In general, we observed a pronounced variation for all platforms except Diagenode. Strikings differences in theoretical versus empirical coverage were observed, especially for Agilent’s platform, reflecting lower uniformity of coverage, with many CpGs not reaching the 10× threshold. Overall, RRBS platforms covered the highest number of CpGs located in CGIs and shores, whereas shelves and open-sea regions were best represented by Roche. Illumina interrogated over twice as many CpGs located in FANTOM5 enhancer regions compared to other platforms. Roche showed the highest coverage of promoters, exons, 5’ untranslated regions (UTRs) and 3’ UTRs, whereas RRBS platforms targeted the most CpGs in introns. When looking at regulatory features defined by ChromHMM14 where promoters were stratified by their activity, HC methods covered more CpGs in active promoter regions, whereas RRBS methods had higher coverage of weak promoters. There were no significant differences between platforms in the coverage of either strong or weak enhancers, whereas Illumina covered the most CpGs located in insulator regions.

Intra-platform reproducibility and inter-platform concordance.

To estimate the level of confidence in obtained results for any given biological experiment and to set appropriate cutoffs for calling differential DNA methylation, it is important to know the approximate level of technical variation in the data. To assess platform reproducibility, we compared DNA methylation data from samples for which the same DNA was used to generate duplicate sequencing libraries (Ref.gDNA, Coriell and HeLa). DNA methylation levels were calculated as the fraction of methylated CpGs over the sum of methylated and unmethylated CpGs. We restricted the analysis to CpG sites covered ≥10× in both samples to calculate pairwise Pearson correlation coefficients.

All platforms showed a very high overall correlation in DNA methylation levels of replicate samples, with Diagenode slightly underperforming (Fig. 4a and Supplementary Fig. 3). The effect of DNA input was estimated using Ref.gDNA libraries prepared with 1,000 ng of DNA for HC or 100 ng input for RRBS methods compared to 500 ng common to all platforms. All platforms, apart from Diagenode, showed very high intra-platform and inter-platform correlation in DNA methylation levels (>0.98) for different DNA inputs (Supplementary Fig. 4).

We next looked at the between-platform concordance in methylation levels for each sample in a pairwise fashion over commonly covered CpGs (94,000–450,000) to enable direct comparison (Supplementary Fig. 5). Methylation levels for each sample per platform were compared to all other platforms to calculate the average correlation over common CpGs (Fig. 4b). Roche HeLa-1 replicate was excluded from the analysis because the hybridization capture failed. We observed a high correlation in called meth-

---

Table 1 | Overview of platform design differences

| Manufacturer | Agilent | Roche | Illumina | Diagenode | NuGen |
|--------------|--------|-------|----------|-----------|-------|
| Platform     | SureSelect Methyl-Seq | NimbleGen SeqCap | TruSeq Methyl Capture EPIC | RRBS Premium | Ovation RRBS Methyl-Seq |
| Enrichment method | In solution hybridization, 150mer biotinylated cRNA baits | In solution hybridization, 60-90mer biotinylated DNA bait | In solution hybridization, biotinylated DNA bait | CCGG restriction enzyme site enrichment | CCGG restriction enzyme site enrichment |
| Targeted region | 84 Mb | 80.5 Mb | 107 Mb | 95.3 Mb** | 95.3 Mb** |
| Number of targeted CpG sites | 3,153,816 | 2,806,466 | 3,346,505 | 3–5 million (4,070,782**) | 3–5 million (4,070,782**) |
| Recommended DNA input | 1 µg | 1 µg | 500 ng | 100 ng | 100 ng |
| Fragmentation method | Ultrasonication | Ultrasonication | Ultrasonication | Restriction enzyme digestion (MSpI) | Restriction enzyme digestion (MSpI) |
| Time-to-result* | 3 days | 5 days | 2 days | 3–5 days | 1 day |
| PCR cycles | 14 | 12 pre- + 14 post-hybridization | 11 | 6-12 (determined by qPCR for each sample) | 12 |
| Multiplexing | 16–96 indexes | 24 indexes | 24 indexes | 24 indexes | 16–96 indexes |
| Advantages | Defined capture regions | Defined capture regions, targets both strands, post-BC capture | Defined capture regions, sample pooling, lower per-sample DNA input requirements | Sample pooling, reduced cost per sample and handling, targets both strands | Short protocol, UMIs for precise duplicate removal, targets both strands |
| Disadvantages | High per-sample DNA input requirements, single-strand capture | High per-sample DNA input requirements, long protocol | Sample pooling, reduced flexibility for sequencing, single-strand capture | Non-specific CpG enrichment, sample pooling, reduced flexibility for sequencing, no duplicate removal | Non-specific CpG enrichment |
| Cost per sample*** | $5 | $5–$50 | $5 | $ | $ |

* per user manual; ** in silico estimate; *** cost per sample: $, <$100; $5, $100–500; $55, >$500
Fig. 2 | Sequencing performance by the platform. a. Pirate plots showing variability in sequencing output by each platform, showing the total number of PF reads (top left), alignment rate (bottom left), the total number of uniquely mapped PF reads (top middle), percent duplicated reads (bottom middle), the total number of de-duplicated uniquely mapped PF reads (top right) and the efficiency of sequencing in terms of the proportion of usable reads (de-duplicated uniquely mapped PF reads) over the total number of PF reads (bottom right). Boxes show median ± range (whiskers) and interquartile range (IQR; boxes). Violin shows distribution density with individual data points per platform (Agilent n = 15, Illumina n = 15, Roche n = 15, Diagenode n = 16 and NuGen n = 16 independent experiments). b–e. Probe hybridization target capture performance by platform (Agilent n = 15, Illumina n = 15 and Roche n = 15 independent experiments) measured as the percentage of on-near bait bases that are on as opposed to near, off-target as the percentage of aligned PF bases that mapped neither on nor near a bait (b), fold enrichment by which the baited region has been amplified above genomic background (c), the mean coverage of targets (d) and the uniformity (e) as the fold over-coverage necessary to raise 80% of target bases to the mean coverage level in those targets. Height of the bar shows the mean ± s.d. (whiskers). f. On-target target capture efficiency variability by platform (Agilent n = 15, Illumina n = 15 and Roche n = 15 independent experiments) represented as the mean ± d. of the fraction of targets covered at a specific depth.

commentary:

TBS comparison to WGBS and Oxford Nanopore methylome. To examine how well each of the more commonly used TBS platforms compares to WGBS as the current gold standard method for near-complete characterization of DNA methylome at single-bp resolution, and Oxford Nanopore as an emerging third-generation sequencing technology that allows for a direct detection of modified bases\(^1\) and complete methylomes\(^2\), we compared TBS to whole methylome datasets for Ref.gDNA and Coriell-NA12878 (Fig. 4d). For Ref.gDNA sequenced by WGBS covering 18,505,516 CpGs (25X), all TBS platforms showed high levels of concordance (\(\rho = 0.98\)), with Diagenode slightly underperforming (\(\rho = 0.94\)). Coriell-NA12878 was represented by two WGBS datasets—from ENCODE (WGBS-EC) covering 24,074,147 CpGs at 28X and from Illumina (WGBS-IL) with 25,093,057 CpGs at 53X and newly generated Nanopore covering 28,230,385 CpGs at 49X. Although we found that Nanopore and Illumina WGBS were interoperable

\(^1\)DNAm-5pct, DNAm-10pct and ZYMO fully methylated control. To examine this further, we compared the concordance between platforms for overall DNA methylation levels on a set of DNA methylation standards generated by mixing DNA from fully unmethylated (ZYMO-UM) and fully methylated controls (ZYMO-FM) at known ratios (0% ZYMO-FM, 5% ZYMO-FM, 10% ZYMO-FM and 100% ZYMO-FM). ZYMO-UM control DNA is isolated from DNMT1/DNMT3b knockout cell line known to have less than 5% methylated DNA\(^1\). All technologies appeared to slightly overestimate methylation levels for ZYMO-UM, and all apart from NuGen tended to underestimate methylation levels of ZYMO-FM (Fig. 4c). This was most pronounced for Diagenode’s platform with 83% estimated methylation for ZYMO-FM, most likely due to the lack of a de-duplication step that removes PCR duplicate reads, driving down the inter-platform concordance for these samples.
with very high degree of overlap (~24 million CpGs) and methylation level concordance (ρ ~0.92) (Extended Data Fig. 3), all TBS platforms displayed, on average, a similar correlation to WGBS datasets (0.96 ± 0.02) and to Nanopore data (0.94 ± 0.02), albeit with the higher CpG overlap with the latter (~2.9 million versus ~3.4 million).

**Calling differentially methylated sites and regions.** Ultimately, many studies aim to identify differentially methylated CpG (DMC) sites or DMRs between samples. To assess the between-platform concordance for calling differential methylation, we compared two pairs of bladder cancer cell lines—one pair with a different genotype and diverse phenotypes (253J and T24) and a second pair of isogenic cell lines with divergent sensitivity to cisplatin (RT112 and RT112-CP). Because the observed technical variability was as high as 5% for some platforms, we set the threshold to >10% methylation difference and q ≤ 0.05 to call DMCs and DMRs.

The number of DMCs depends on the number of CpGs assayed, their location with respect to genomic features and inherent technical biases for each of the platforms. To calculate DMCs, a set
of overlapping CpGs within the pair is analyzed per platform but different sets of CpGs between platforms as a consequence of differences in platform CpG coverage. As expected, the bulk of identified DMCs for all platforms showed larger methylation differences for the genetically distinct pair of cell lines compared to the isogenic RT112 versus RT112-CP pair, with only a small subset of DMCs commonly identified by all five platforms as a consequence of differences in platform CpG coverage. Furthermore, the effect of low overlap in the CpG sites selected is of particular concern as downstream analyses regularly require subsetting to a set of CpG sites that are present in all datasets. However, it came at the cost of reduced DNA methylation concordance from an average of 0.97 to 0.80 after imputation (Fig. 5d and Supplementary Fig. 7). Imputation can be used to estimate the methylation value for each missing CpG site.\textsuperscript{18,19} and, thereby, can be used to help increase platforms’ intra-operability and inter-operability. Samples with duplicate sequencing libraries, one clonal (Coriell-NA12878) and one heterogeneous admixture of blood cell types (Ref.gDNA), were further analyzed to measure the effect of imputation on CpG methylation concordance and overlap. We used BoostMe\textsuperscript{19} to impute the vast majority of missing CpGs, leveraging information learned from two neighboring CpGs of the same dataset, with or without a distance threshold (Fig. 5e–f). Imputation without distance threshold increased the overlap between platforms almost ten-fold, from 0.79 million (10.35%) to 7.6 million (97%) CpG sites (Fig. 5c, Extended Data Fig. 4 and Supplementary Fig. 6). However, it came at the cost of reduced DNA methylation concordance from an average of 0.97 to 0.80 after imputation (Fig. 5d and Supplementary Fig. 7). Mean absolute error (MAE), which estimates by how much each methylation value is off, increased from an average of 0.05 to 0.14 for a given sample pair.

**Imputation to improve platform interoperability.** The platform similarity as a function of the number of overlapping CpG sites is of particular concern as downstream analyses regularly require subsetting to a set of CpG sites that are present in all datasets. Furthermore, the effect of low overlap in the CpG sites selected is further exacerbated as the number of samples/datasets increases. Imputation can be used to estimate the methylation value for each missing CpG site\textsuperscript{18,19} and, thereby, can be used to help increase platforms’ intra-operability and inter-operability. Samples with duplicate sequencing libraries, one clonal (Coriell-NA12878) and one heterogeneous admixture of blood cell types (Ref.gDNA), were further analyzed to measure the effect of imputation on CpG methylation concordance and overlap. We used BoostMe\textsuperscript{19} to impute the vast majority of missing CpGs, leveraging information learned from two neighboring CpGs of the same dataset, with or without a distance threshold (Fig. 5e–f). Imputation without distance threshold increased the overlap between platforms almost ten-fold, from 0.79 million (10.35%) to 7.6 million (97%) CpG sites (Fig. 5c, Extended Data Fig. 4 and Supplementary Fig. 6). However, it came at the cost of reduced DNA methylation concordance from an average of 0.97 to 0.80 after imputation (Fig. 5d and Supplementary Fig. 7). Mean absolute error (MAE), which estimates by how much each methylation value is off, increased from an average of 0.05 to 0.14 (Fig. 5e and Supplementary Fig. 8), and root mean square error (RMSE) increased from 0.09 to 0.20 (Fig. 5f and Supplementary Fig. 9), whereas MAE corrected for the number of CpGs present in the
Fig. 5 | Differential methylation calls by the platform and imputation. a,b. Concordance between DMC sites (a) and DMRs (b) as 1,000-bp windows by each platform, in genetically distinct pairs of cell lines T24 versus 253J (left) and a pair of isogenic cell lines RT112 versus RT112–CP (right). Top, UpSet plot shows the number of common calls for each combination; middle violin plot shows distribution density of significant methylation differences (>10%); and bottom box plot shows the mean with range (whiskers) and IQR (boxes) of significant q values (q < 0.05) identified by each platform for a given DMC/DMR. c, Bar plot showing the number of common CpGs covered by all five platforms for Coriell-NA12878 (n = 10) and Ref.gDNA (n = 10), before and after imputation of missing CpG values using two neighboring CpGs applying different conditions (i, no distance threshold; i_1,000, 1,000-bp distance threshold; and i_25, 25-bp distance threshold). d–f, Height of the bar shows the mean ± s.d. (whiskers) of average Pearson correlation coefficient (d), average MAE (e) and average RMSE (f) of n = 90 individual pairwise comparisons for Coriell-NA12878 (n = 10, average mean coverage 42 ± 11) and Ref.gDNA (n = 10, average mean coverage 51 ± 20) samples analyzed by the five TBS platforms.

dataset decreased from an average of 0.12 to 0.002 (Supplementary Fig. 10). When a distance threshold of <25 bp was applied to impute neighboring CpGs, the total number of overlapping CpGs was 2.5 million (32%), whereas the intra-platform concordance after imputation remained high (ρ = 0.89, MAE = 0.10).

Discussion

To answer important biological questions in the epigenetics of health and disease, we must understand the possibilities and limitations of the tools that we use. Over the past decade, many studies have leveraged the power of bisulfite sequencing methods to advance knowledge on epigenetics of non-communicable diseases, to identify clinical biomarkers and to understand normal development20–24. Several companies offer off-the-shelf pre-designed products for targeted analysis of DNA methylation at the genome-wide level. We compared five commercial platforms (Agilent SureSelect Methyl-Seq, Roche NimbleGen SeqCap EpiGiant, Illumina TruSeq Methyl Capture EPIC, Diagenode Premium RRBS and NuGen Ovation RRBS Methyl-Seq) and examined their performance, similarities and differences. To mimic a real-life experimental setup, for each platform the same panel of 11 samples was processed, generating 16 libraries, including technical replicates, sequenced to a similar number of usable reads and analyzed using the same bioinformatics pipeline.

The experiment protocol complexity, length and DNA input requirements often dictate the choice of the platform. Small-scale studies focused on in vitro experiments using cell lines or animal models have fewer constraints than high-throughput applications.
using scarce clinical samples. NuGen RRBS offers the shortest and least complex protocol that requires only 100 ng of DNA, suitable for high-throughput applications using scarce clinical samples. On the other hand, smaller-scale studies focused on in vitro experiments using cell lines, animal models or abundant clinical specimens have fewer constraints, allowing the use of HC platforms. Among HC platforms, Illumina’s protocol can be completed in 2 days and requires half the amount of DNA per sample (500 ng) by multiplexing four samples, albeit reducing the flexibility in experimental design.

Differences in design, size of the targeted region and experimental strategies for library preparation and target capture translate into stark differences in the cost-effectiveness of sequencing. To achieve a similar number of uniquely mapped de-duplicated reads, HC platforms required ~20% more sequencing than RRBS platforms. In addition to RRBS platforms, only the Roche HC platform targets both strands at the same ratio, whereas Illumina covers both strands only for common single-nucleotide polymorphisms (SNPs), permitting the distinction between a C-to-T SNP and a methylated CpG site.

Estimates of 5-methylcytosine (5mC) levels may be biased by experimental and bioinformatic factors. Our results reveal high within-platform and between-platform concordance in measured 5mC levels of the overlapping CpG sites. Although all platforms slightly underestimated global 5mC levels for fully methylated DNA standards, this was particularly pronounced for Diagenode’s platform, reflected by the lower between-platform concordance of methylated ZYMO controls. Because the sample source and sequencing conditions were the same for all, the bias must have been introduced during the library preparation and/or data de-duplication. Olova et al.\(^\text{26}\) found that the BC is the main driver of sequencing biases that are further amplified by PCR amplification. Although all platforms perform adaptor tagging before BC, whereas Roche performs hybridization capture before BC,\(^\text{26}\) Roche and Illumina platforms perform hybridization capture before BC, whereas Roche performs hybridization capture after BC.\(^\text{26}\) Although the data processing of HC platforms included a de-duplication step based on read coordinates, only NuGen’s protocol incorporates UMIs, enabling precise read de-duplication to account for bias introduced by PCR amplification. In contrast, Diagenode’s platform does not allow for PCR de-duplication, rendering it susceptible to biased estimation of 5mC levels. Conversion efficiency was above 99% for all platforms except for Roche, where the rate was over 98%. Over-conversion may lead to underestimation of 5mC levels; however, fully methylated spike-in was not included in all platforms, so we could not compare the over-conversion rates.

The choice of method defines the space for the discovery of CpG loci or regions associated with a specific phenotype. Each of the platforms targeted different regions of the genome, and the overlap between CpG loci covered by all was modest. Overall, RRBS methods covered more CpGs ≥10x, which were not targeted by HC platforms. We provide a breakdown of genomic feature coverage by platform that can be used as a guide for platform selection depending on the biological question. Calling of DMCs is limited on the CpGs covered in common between the samples compared. Given a pronounced variation in the number of covered CpGs within each platform, the larger the number of samples, the lower the overlap. To fully capitalize on a dataset, two strategies can be employed. First, methylation values can be aggregated in defined regions (for example, promoters) or tiling regions (for example, 1,000-bp windows) to call DMRs. We found that, although many of the DMCs called by the five platforms did not overlap, most DMRs were commonly identified by all platforms. Alternatively, to take the most advantage of the base-level resolution offered by TBS, missing CpGs can be imputed. We showed that imputation of missing CpG sites is a viable option to increase inter-platform operability while maintaining high concordance for downstream analysis. Although intra-sample heterogeneity and discordant methylation may pose a challenge for precise imputation, by selecting the stringency of the distance threshold and associated MAE, users can determine the appropriate methylation difference cutoff for calling DMCs.

In summary, our study provides an in-depth analysis of the TBS platforms’ comparative performance and characteristics, allowing users to determine the optimal technology for methylome analysis depending on their needs and restrictions. It also provides guidance for cross-platform data integration.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01336-9.

Received: 6 September 2021; Accepted: 28 April 2022; Published online: 2 June 2022

**References**

1. Schubeler, D. Function and information content of DNA methylation. Nature 517, 321–326 (2015).
2. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. Nat. Rev. Genet. 11, 191–203 (2010).
3. Stitz, S., Caberley, P. C., Statham, A. L. & Clark, S. J. Mining cancer methylomes: prospects and challenges. Trends Genet. 30, 75–84 (2014).
4. Gu, H. et al. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. Nat. Methods 7, 133–136 (2010).
5. Guo, S. et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat. Genet. 49, 635–642 (2017).
6. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189 (2009).
7. Meissner, A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 33, 5868–5877 (2005).
8. Kacmaryrzek, T. J. et al. ‘Same difference’: comprehensive evaluation of four DNA methylation measurement platforms. Epigenetics Chromatin 11, 21 (2018).
9. Warncke, P. M. et al. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res. 25, 4422–4426 (1997).
10. Woyaca, T. K., Borgbo, T. & Hansen, L. L. Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics 4, 231–234 (2009).
11. Ebbert, M. T. et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics 17, 239 (2016).
12. Kirvija, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9, 72–74 (2011).
13. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
14. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216 (2012).
15. Rhee, I. et al. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. Nature 416, 552–556 (2002).
16. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods 14, 407–410 (2017).
17. Gershman, A. et al. Epigenetic patterns in a complete human genome. Science 376, eaat5809 (2022).
18. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 18, 67 (2017).
19. Zou, L. S. et al. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. BMC Genomics 19, 390 (2018).
20. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–322 (2009).
21. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann. Oncol. 31, 745–759 (2020).
22. Landau, D. A. et al. Locally disordered methylation forms the basis of chromatin in chronic lymphocytic leukemia. Cancer Cell 26, 813–825 (2014).
23. Li, S. et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat. Med. 22, 792–799 (2016).
24. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. Nature 567, 479–485 (2019).
25. Olova, N. et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 19, 33 (2018).
26. Li, Q. et al. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. Nucleic Acids Res. 43, e81 (2015).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022
Methods

Sample origin and processing. The reference gDNA sample (Ref.gDNA) was obtained by pooling DNA extracted from peripheral whole blood cells of 20 adult healthy volunteers who provided informed consent under 15/YH/0311 of the UCL Biobank Ethics Committee. Blood samples were collected in EDTA-treated 10-ml BD Vacutainer tubes and centrifuged at 1,900g for 10 minutes at 4°C using a refrigerated centrifuge to separate blood cells from plasma. Buffy coat (0.5 ml) was transferred using a Pasteur b) to a clean Falcon tube containing 4.5 ml of HEMAgene BUFFY COAT DNA staining reagent (Oargoene), and the DNA was extracted using QIAamp DNA Mini Kit (Qiagen). The Human Methylated & Non-methylated DNA Set (ZYMO) was used to generate DNA methylation standards: fully methylated (ZYMO-FM), 10% and 5% methylated (DNAM-10pct and DNAm-5pct) and unmethylated (ZYMO-UM). The negative control (ZYMO-UM) DNA in the set comes from a human HTCL116 DKO non-methylated DNA that contains genetic knockouts of both DNA methyltransferases DNM1T1 (→) and DNM1T3 (→) and has low levels of DNA methylation, whereas the fully methylated DNA control was generated by the manufacturer by treating the HTCL116 DKO with M. SssI methyltransferase. Sample quality was determined using an Agilent 2100 Bioanalyzer 1000 DNA chip to assess the DNA integrity, and sample purity was estimated using a NanoDrop spectrophotometer. Quantification was performed using Qubit dsDNA BR Assay (Invitrogen).

Library preparation and target enrichment with Agilent SureSelect Methyl-Seq kit. To generate SureSelect Methyl-Seq sequencing libraries (Agilent), we followed the manufacturer’s protocol. In brief, we sheared 1 μg of the gDNA on Covaris S2 ultrasonicator to obtain ~150-bp fragments. Fragmented DNA was cleaned up using AMPure XP beads (Beckman Coulter), followed by end-repair, A-tailing, second cleanup and adapter ligation before the final cleanup. Quality and quantity of the libraries were checked using a Bioanalyzer dsDNA High Sensitivity Chip (Agilent) and Qubit HS dsDNA Assay (Life Technologies, Thermo Fisher Scientific) before overnight hybridization (minimum 16 hours) with biotinylated oligo RNA baits, followed by streptavidin-conjugated magnetic bead pulldown and wash steps. The captured library was then bisulfite-converted using EZ DNA Methylation-Gold Kit (ZYMO), followed by eight cycles of PCR amplification and AMPure XP bead cleanup. Second PCR amplification with six cycles of the bisulfite-converted libraries was performed to introduce sample indices, followed by final AMPure XP bead cleanup. Library concentration was measured using Qubit dsDNA HS kit, and the library size distribution was checked on a Bioanalyzer High Sensitivity DNA chip.

Library preparation and target enrichment with Roche NimbleGen SeqCap EpiGiant kit. NimbleGen SeqCap EpiGiant (Roche) libraries were prepared according to manufacturer instructions, as follows: 1 μg of the gDNA sample was sheared using Covaris S2 ultrasonicator to an average DNA fragment size of 200 bp; sheared DNA was end-repaired and cleaned up with magnetic beads, followed by A-tailing of the 3’ end, second cleanup, adapter ligation and double SPRI cleanup. Libraries were bisulfite-converted using EZ DNA Methylation Lightning Kit (ZYMO), followed by 12 cycles of PCR amplification using Kapa HiFi Uracil Polymerase (Roche) and purified with AMPure XP magnetic beads. Library concentration was measured using the Qubit dsDNA HS kit; purity was measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific); and the library size distribution was checked on a Bioanalyzer High Sensitivity DNA chip.

Library preparation and target enrichment with Illumina TruSeq Methyl Capture Epic kit. The TruSeq Methyl Capture Epic kit was a gift from Illumina. Sequencing libraries were prepared according to the manufacturer’s instructions. In brief, 500 ng of each gDNA sample was sheared on a Covaris S2 ultrasonicator to a median size of 160 bp, followed by a cleanup step using AMPure XP magnetic beads, end-repair reaction, second magnetic bead cleanup, A-tailing, adaptor ligation and final cleanup. Four samples containing adapters were pooled together to a total mass of 2 μg and hybridized with a biotinylated DNA capture panel for 35 minutes at 58°C. Target DNA fragments were captured by streptavidin-conjugated magnetic beads and washed, followed by a second overnight hybridization (touchdown from 95°C to 58°C), streptavidin-conjugated magnetic bead pulldown and wash step. Enriched libraries were bisulfite-converted using the EZ DNA Methylation Lightning kit with magnetic bead de-sulphonation and cleanup. Eleven cycles of PCR amplification were performed using Kapa HiFi Uracil+ Polymerase, available separately from the kit. Final libraries were bead-purified to remove adapters and quantified using the Qubit dsDNA HS kit, and the library size distribution was checked on a Bioanalyzer High Sensitivity DNA chip.

RRBS library preparation using Diagenode's RRBS Premium kit. The protocol for Premium RRBS kit (Diagenode) library preparation included enzymatic digestion of 100 ng of gDNA with MspI, followed by end-repair, adapter ligation and size selection by bead purification. Each library was quantified by quantitative PCR (qPCR) to determine library concentration, and eight samples were pooled in equimolar amounts determined by an Excel pooling protocol provided by the manufacturer. Pooled libraries were bisulfite-converted using manufacturer-provided BS Conversion Reagent and de-sulphonated on columns. The second qPCR was performed to determine the optimal number of cycles for the final amplification step for each pool (11 and 12 cycles) before final cleanup. Library concentration was measured using Agilent’s nCode DNA High Sensitivity HS kit, and the library size distribution was checked on a Bioanalyzer High Sensitivity DNA chip.

RRBS library preparation using NuGen's Ovation RRBS Methyl-Seq System 1–16. The Ovation RRBS Methyl-Seq System 1–16 (NuGen, now Tecan) user manual was followed to generate sequencing libraries by enzymatically digesting 100 ng of gDNA using MspI, followed by end-repair, adapter ligation and a final repair step. Generated libraries were bisulfite-converted using EpTect Fast DNA Bisulfite Kit (Qiagen) purchased separately from the kit. Converted libraries were amplified by PCR using 12 cycles and purified using Agencourt RNAClean XP Beads and fragment distribution was checked on a Bioanalyzer High Sensitivity DNA chip. An RRBS design BED file was created by in silico cutting of hg19 sequence at CG within CCGG motif and filtering to >100-bp length ranges and 100-bp window from each end.

Sequencing on an Illumina HiSeq 2500 instrument. All generated, targeted bisulfite-converted libraries were sequenced on the HiSeq2500 using HiSeq v4 SBS reagents (Illumina) at 2× 100 bp. Barcoded samples per each library were multiplexed, and libraries were denatured with NaOH and diluted to a final concentration of 2 nM. Denatured libraries were loaded onto cBot for cluster generation on separate flowcell lanes in HighOutput or RapidRun mode according to the manufacturer’s protocol. Loading concentrations and percent of PhiX spike-in libraries for each platform are as follows: Agilent at 15±4 pM with 5% of equimolar PhiX, Roche at 16.5±3 pM with 5% equimolar PhiX, Illumina at 11±3 pM with 10% equimolar PhiX, Diagenode at 16±3 pM with 5% of 12-pM PhiX and NuGen at 11.5±3 pM with 5% of equimolar PhiX. Sample pools were sequenced over several lanes/runs, and raw FASTQ reads were merged before the analysis.

TBS data analysis. Raw sequencing data processing. For all targeted bisulfite sequencing platforms, Bismark² data analysis pipeline was applied with few platform-specific modifications. Samples were demultiplexed either onboard Illumina HiSeq2500 or offline using bcld2fastq Conversion Software version 1.8.4 (Illumina). The quality of raw reads in FASTQ format was checked using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) version 0.11.9 (Supplementary Fig. 1). For each sample platform, FASTQ files from different sequencing runs were merged before alignment. NuGen FASTQ reads contain an additional 1–3 inserted nucleotides and were trimmed using a diversity-trimming Python script provided by the manufacturer (https://github.com/NuGeneTechnologies/NuMetRRBS). Any contaminating adapter sequence was removed using TrimGalore (https://github.com/FelixKrueger/TrimGalore/TrimGalore) version 0.6.5 in paired-end mode with default conditions for adapters and trim1 option for Agilent, Roche and Illumina, whereas, for Diagenode’s platform, –rrbs option was used, and, for NuGen, read 2 adapter was specified to generate sequencing library by enzymatically digesting the human genome build hg19 in paired-end mode using Bowtie2 version 2.3.4.2 and default settings. After alignment, PCR duplicates in Agilent, Roche and Illumina were marked and discarded using deduplicate_bismark perl script. For RRBS data, the second step of de-duplication is not recommended as all reads start with the same sequence. However, the NuGen platform contains UMIs that enable for precise removal of duplicates using Duplicate Matching Python script provided by the manufacturer (https://github.com/NuGeneTechnologies/NuMetRRBS). Bismark’s module bismark_methylation_extractor with options --ignore_r2 2, --no_overlap, --compressive, --merge_nonCpG, --bedGraph and --gzip was used to generate output files with methylation status for each individual CpG, coverage and genomic coordinates. Bismark’s coverage2tiosine tool with and without --merge_CpG option was used to obtain methylation status for each CpG in the genome, both strands, along with trimucleotide context, coverage, and methylation percent. The efficiency of bisulfite conversion was verified by the ratio of C-to-T conversion of CHG and CGH (non-CG) dinucleotides. MultiQC³ version 1.10.1 was used for visualizing aggregated quality metrics.

DNA methylation data analysis. Downstream analysis was performed in Unix (RHE) and R environment (version 3.4.1). We did not use Repeat masking for any of the analyses. For hybridization capture methods, the Picard
Imputation. We used BoostMe to impute the vast majority (that is, 99.95%) of missing CpG sites, leveraging information learned from just the neighboring CpG within the same dataset. BoostMe was run with a 60:20:20 split among the training, validation and test datasets. On average, the training dataset was formed from 2 million randomly selected CpG sites, whereas the validation and test datasets were made up of 0.67 million randomly selected CpG sites. Imputation accuracy was 92.1% and 0.971 area under the receiver operating characteristic (AUROC) in cross-validation analysis, and 92.1% accuracy and 0.971 AUROC in the test dataset. Imputation accuracy can be difficult in allosomes as the methylation composition can be considerably different from autosomal chromosomal homoeologies to the samples’ sex. As such, after excluding the allosomes, there is an average of 7.59 million CpG sites, of which an average of 4.23 million CpG sites are missing in each dataset. BoostMe achieves a 92.1% accuracy and 0.971 AUROC in cross-validation analysis and achieves 92.1% accuracy and 0.971 AUROC with the test dataset. The larger dataset not used for training or cross-validation. In this manner, by imputing missing CpG sites with BoostMe, the overlap among the five platforms increases from 10.35% (0.79 million CpG sites) to 97% (7.6 million CpG sites).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. The datasets generated and analyzed in the current study, including all raw targeted bisulfite sequencing, WGBS of Ref.gDNA and Nanopore sequencing data, have been deposited in the European Nucleotide Archive repository under accession number PRJEB65109 and are freely available. Raw WGBS sequencing data for the Coriell-NA12878 WGBS_JC sample generated by the ENCODE project Consortium was downloaded from the ENCODE project (experiment: ENCSR890UQJ, library: ENCLB898WPJ) (https://www.encodeproject.org/experiments/ENCSR890UQJ), and CpG count files for WGBS_IL sample were downloaded from Illumina BaseSpace Hub (https://basespace.illumina.com/dataset) under sample name WGBS_P3 from HiSeq 4000: TruSeq DNA Methylation (NA12878, 2×76 bp).

Code availability. The code used for annotation, differential methylation analysis, plotting and imputation is available in the GitHub repository at https://github.com/ucl-medical-genomics/EpiCapture.

References. 27. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulflte-sequencing data. Bioinformatics 27, 1571–1572 (2011). 28. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016). 29. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.12.11–11.12.34 (2014). 30. Gu, Z., Ellis, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849 (2016). 31. Zhu, L. et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 11, 237 (2010). 32. Akalin, A. et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 13, R87 (2012). 33. Wang, H. Q., Tuominen, L. K. & Tsai, C. J. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. Bioinformatics 27, 225–231 (2011). 34. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. Bioinformatics 33, 2381–2383 (2017). 35. Lawrence, M. et al. Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9, e1003118 (2013). 36. Morgan, M. & Shepherd, L. AnnotationHub: client to access AnnotationHub resources. R package version 3.2.0. https://bioconductor.org/packages/release/bioc/html/AnnotationHub.html (2022). 37. Lawrence, M. HelloRanges: introduce *Ranges to bedtools users. R package version 1.2.0. https://bioconductor.org/packages/release/bioc/html/HelloRanges.html (2022). 38. Khan, A. & Mathelier, A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. BMC Bioinformatics 18, 287 (2017). 39. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
and Technological Development of Serbia (2011-2019/III-41026 and 451-03-68/2020-14/200043); and the Science Fund of the Republic of Serbia (PROMIS/2020/6060876). I.M. is supported by the Biotechnology and Biological Sciences Research Council (grant no. BB/M009513/1). S.B. has received funding from the Wellcome Trust (218274/Z/19/Z) and a Royal Society Wolfson Research Merit Award (WM100023). A.F. received support from the UCL/UCLH Biomedical Research Centre, the Medical Research Council (MR/M025411/1), Prostate Cancer UK (MA_TR15_009) and the Biotechnology and Biological Sciences Research Council (BB/R009295/1). S.R. received funding from Orchid. We further acknowledge support from D. Turner and B. Sipos (Oxford Nanopore Technologies) for the generation of the Nanopore sequencing data and from the CRUK–UCL Centre-funded Genomics and Genome Engineering and Bioinformatics Translational Technology Platforms.

Author contributions
M.T., A.F. and S.B. conceived and designed the study. M.T. and S.R. performed the hybridization capture and RRBS experiments. P.D. and H.V. sequenced the libraries. M.T. and J.B. processed raw sequencing data. M.T. performed analysis of TBS data. I.M. analyzed WGRS and Nanopore data and performed imputation analysis. M.T., A.F. and S.B. interpreted the results. M.T., A.F. and S.B. wrote the manuscript. All authors read and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01336-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01336-9.

Correspondence and requests for materials should be addressed to Miljana Tanić or Stephan Beck.

Peer review information Nature Biotechnology thanks Miguel Branco, Alexander Dobrovic and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Sequencing data processing quality metrics produced by MultiQC. 

a Bismark alignment rates for uniquely, ambiguously, or unaligned reads for each sample by platform; 
b Percent of reads aligning to top or bottom DNA strand for each sample by the platform; 
c Global methylation levels of CpG dinucleotides for each sample by the platform; 
d The global cytosine methylation level in CHG context for each sample by the platform used an estimate of sodium bisulfite under-conversion rates; 
e The global cytosine methylation level in CHH context for each sample by the platform used an estimate of sodium bisulfite under-conversion rates. 
f-g M-bias plot shows the average percentage methylation and coverage across read length for each sample. Each line represents a sample. Methylation bias for the forward sequencing read by platform (f); Methylation bias for the reverse sequencing read by platform (g).
Extended Data Fig. 2 | Target depth of coverage. The fraction of targets covered at specific depth of sequencing for each sample by the platform: Agilent (a), Illumina (b) and Roche (c). Each sample is represented by a line.
Extended Data Fig. 3 | Intra-platform concordance. Scatterplot showing pairwise Pearson correlation coefficient for Coriell NA12878 data from, WGBS EC vs. WGBS IL (a), Nanopore vs. WGBS IL (b), and Nanopore vs. WGBS EC (c).
Extended Data Fig. 4 | Platform interoperability. Interoperability between platforms for Coriell NA12878 (left) and Ref.gDNA (right) before imputation (first row), after imputation without distance threshold (second row), after imputation with 1000 bp distance threshold (third row) and after imputation with 25 bp distance threshold (fourth row). Venn diagram showing CpGs overlapping between the platforms.
Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - *Give P values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

*Our web collection on statistics for biologists contains articles on many of the points above.*

Software and code

Policy information about availability of computer code

Data collection | The demultiplexed FASTQ files from the sequencing center were shared with us via FTP, or downloaded using Illumina BaseSpace manager.

Data analysis

Samples were demultiplexed either onboard Illumina HiSeq2500 or offline using bcl2fastq Conversion Software v1.8.4 (Illumina) Sequencing output was quality checked with FastQC v0.11.9 (Babraham Institute, https://www.babraham.ac.uk/) and adapter sequences were trimmed with Trimgalore version 0.6.5, and reads are aligned to the UCSC hg19 reference assembly using Bismark v0.10.1 (Krueger and Andrews, 2011). Read deduplication for NuGen was carried out using NuDup, (NuGEN, https://github.com/nugentechologies/nudup). For hybridization capture methods PICARD CollectIoMetrics tool v2.3.0 was used to obtain statistics on HC efficiency. BEDTools v2.25.0 and custom R, sed and awk scripts were used to calculate and visualize the target region coverage as a function of read depth. MultiQC v1.10.1 was used for visualizing aggregated quality metrics. R version 3.4.1 was used with following packages: ‘plyr’ [1.4.0], ‘stringr’ [1.2.0], ‘data.table’ [1.9.6]. To compare target breadth of coverage UpSet plot from ‘ComplexHeatmap’ [2.10.0] and Venn diagram from ‘ChipPeakAnno’ [3.28.1] were used to visualize the intersection of genomic ranges and CpG sites. Feature annotation was done using ‘annotatr’ R package, and analyzed with ‘GenomicRanges’ [1.46.1] and ‘GenomicFeatures’ [1.46.5]. Bismark produced strand-specific methylation calls were analyzed with ‘MethylKit’ [1.2.0]. The genomic annotations are determined by functions from ‘GenomicFeatures’ [1.18.13] and data from the ’TrxDb’ [3.2.2] and ‘org.hg19.db’ packages. Jaccard similarity index was calculated using functions from ‘HelloRanges’ [1.3.1] Intervene shiny app (https://asntech.shinyapps.io/intervene/) was used to visualize the Jaccard similarity index matrix. BoostMe (https://github.com/luizou/boostme) was used for imputation. The custom code used for annotation, differential methylation analysis, plotting and imputation is available from Github repository at https://github.com/ucl-medical-genomics/EpiCapture.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Portfolio guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy.

The datasets generated and analyzed in the current study including all raw targeted BS-seq, WGBS of Ref.gDNA and Nanopore sequencing data have been deposited in the European Nucleotide Archive repository under accession no. PRJEB46506 (https://www.ebi.ac.uk/ena/browser/view/PRJEB46506?show=reads) and are freely available. Raw WGBS sequencing data for Coriell NA12878 WGBS_EC sample generated by the ENCODE Project Consortium 26 was downloaded from the Encode Project [experiment: ENCSR890UQO, library: ENCLB898WPW] (https://www.encodeproject.org/experiments/ENCSR890UQO/), and CpG count files for WGBS_IL sample were downloaded from Illumina BaseSpace Hub (https://basespace.illumina.com/datacentral) under sample name WGBS_P3 from HiSeq 4000: TruSeq DNA Methylation (NA12878, 2x76) dataset.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-lif.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
For each TBS platform a total of 16 libraries were generated (Fig. 1a) using the same source DNA from 11 biological replicates generating a total of 80 TBS libraries, in addition to n=2 WGBS (Coriell-NA12878 and Ref.gDNA) and n=1 Nanopore (Coriell-NA12878) libraries.

The primary goal of the study was to measure the technical (random) variability and correlation of called DNA methylation levels and other metrics within and between TBS platforms over the same set of samples. To determine the number of samples we followed a statistical rule of thumb based on the fact that the width of the confidence interval (CI) is a function of the square root of the number of observations. The width of a CI involving estimation of variability and sample size decreases rapidly until 12 observations are reached and approaches the asymptote at 15 observations (Gerald van Belle, Statistical Rules of Thumb, 2008). The half-width of the 95% confidence interval at n=15 is 0.56 assuming normal distribution with n-1 degrees of freedom.

Data exclusions
Library prep failed for samples Agilent DNA: 10pct and Illumina DNAQm: 5pct, while Roche Hela 1 library failed target capture and was excluded from downstream analysis.

Replication
For each TBS platform a total of 16 libraries were generated (Figure 1a) using the same source DNA from 11 biological replicates that consisted of: reference gDNA isolated from human peripheral blood cells (Ref.gDNA) at two different DNA inputs (recommended by the manufacturer and 500 ng) in duplicate; Coriell NA12878 and Hela cell lines processed in duplicate; four DNA methylation standards generated from ZYMO fully methylated and unmethylated control samples; a pair of genetically and phenotypically divergent bladder cancer cell lines and a pair of isogenic bladder cancer cell lines with different sensitivity to cisplatin treatment.

All replication attempts were successful, except for Hela sample processed with Roche TBS platform.

Randomization
Not relevant to our study. We used the same source DNA to generate libraries used for comparison, sequenced them on the same instrument (HiSeq 2500) and analyzed using the same bioinformatic pipeline.

Blinding
Not relevant to our study. Blinding was not required since all samples processed by different TBS platforms came from the same source DNA and the data was processed using the same computational pipeline designed to be unbiased in their evaluation of methods.

Reporting for specific materials, systems and methods
We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
Materials & experimental systems

| Involved in the study |
|-----------------------|
| Antibodies            |
| Eukaryotic cell lines |
| Palaeontology and archaeology |
| Animals and other organisms |
| Human research participants |
| Clinical data         |
| Dual use research of concern |

Methods

| Involved in the study |
|-----------------------|
| ChiP-seq              |
| Flow cytometry        |
| MRI-based neuroimaging |

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
Cell lines 253J, T24 and RT112 were obtained from the American Type Culture Collection (#TCP-1020). RT112-CP cell line derived from RT112 by growing the cell under selection for cisplatin resistance was a gift from Prof Jim Catto, Sheffield University. The Corbell-NA12878 DNA sample was purchased from Corbell Cell Repositories. Hela cell genomic DNA was purchased from New England Biolabs. The Human Methylated & Non-methylated DNA Set were purchased from Zymo Research.

Authentication
None of the cell lines were authenticated.

Mycoplasma contamination
Cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register)
None of these misidentified lines were used.

Human research participants

Policy information about studies involving human research participants

Population characteristics
The reference genomic DNA sample (Ref.gDNA) was obtained by pooling DNA extracted from peripheral whole blood cells of six adult healthy volunteers (four male and two female), age range (28-62).

Recruitment
Adult healthy UCL CI staff members were approached to donate blood samples providing informed consent under Ethics approval 15/YH/0311 of the UCL BioBank for Health and Disease. The plasma samples were pooled to ensure anonymity and make a reference genomic DNA sample (Ref.gDNA) representing an admixture of healthy blood plasma samples.

Ethics oversight
Informed consent was obtained under study 15/YH/0311 approved by the UCL BioBank for Health and Disease ethics committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.