Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Systematic Evaluation of Short-Read, Long-Read, and Microarray Platforms for Integrated Genomic and Epigenomic Profiling in Brain Tumors

Jan Bińkowski

jan.binkowski@pum.edu.pl

Pomeranian Medical University

Katarzyna Sokołowska
Pomeranian Medical University

Maria Stępień
Pomeranian Medical University

Kamila Stanisz
Pomeranian Medical University

Kamila Marszałek
Pomeranian Medical University

Michał Kowalski
Pomeranian Medical University

Patrycja Przybyłowicz
Pomeranian Medical University

Thierry Wettering
Pomeranian Medical University

Joanna Dowejko
Pomeranian Medical University

Bartłomiej Grygorcewicz
Pomeranian Medical University

Piotr Czupryński
Pomeranian Medical University

Jakub Baranowski
Pomeranian Medical University

Bartłomiej Pala
Pomeranian Medical University

Paweł Białas
Pomeranian Medical University

Bartłomiej Sagan

Pomeranian Medical University
Leszek Sagan
Pomeranian Medical University
Andrzej Ossowski
Pomeranian Medical University
Tomasz Wojdacz
Pomeranian Medical University

Research Article

Keywords: 5-base, EPICv2, Oxford Nanopore Technologies, MGMT, DNA methylation

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11364944/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. T.K.W. has received honoraria from Illumina. Supporting entities had no influence on the study, the interpretation of the results, or the content of the manuscript, and did not review or approve the manuscript prior to submission.

Systematic Evaluation of Short-Read, Long-Read, and Microarray Platforms for Integrated Genomic and Epigenomic Profiling in Brain Tumors

Jan Bińkowski[1,2*], Katarzyna Sokołowska[1,2], Maria Stępień[3], Kamila Stanisz[2,3], Kamila Marszałek[3], Michał Kowalski[2,3], Patrycja Przybyłowicz[1,2], Thierry van de Wettering[2], Joanna Dowejko[3], Bartłomiej Grygorcewicz[2,3], Piotr Czupryński[2,4], Jakub Baranowski[5], Bartłomiej Pala[5], Paweł Białas[5], Bartłomiej Sagan[5], Leszek Sagan[5], Andrzej Ossowski[2,3], Tomasz K. Wojdacz[1,2,3]

1. Independent Clinical Epigenetics Laboratory, Pomeranian Medical University in Szczecin
2. Regional Center for Digital Medicine, Pomeranian Medical University in Szczecin
3. Department of Genomics and Forensic Genetics, Pomeranian Medical University in Szczecin
4. Clinical Trials Support Center, Pomeranian Medical University in Szczecin
5. Department of Neurosurgery, Pomeranian Medical University in Szczecin

*corresponding author: jan.binkowski@pum.edu.pl

Abstract

Background: Recent advances in sequencing- and microarray-based technologies for genome and epigenome profiling have created an urgent need to assess their analytical performance, technological maturity, and potential clinical utility. We therefore evaluated short-read 5-base sequencing, long-read Oxford Nanopore Technologies (ONT) sequencing, and EPICv2 microarrays using brain tumor samples as a clinically relevant benchmark.

Methods: The study included 5-base sequencing and EPICv2 microarray data generated for 16 brain tumor samples initially diagnosed as glioblastoma (GBM), together with ONT sequencing data available for 10 of 16 samples. Data were processed using technology-specific workflows, including DRAGEN somatic, EPI2ME wf-somatic-variation, and sesame R package for 5-base, ONT and microarray data respectively. The platforms were compared in terms of sequencing yield, basecalling quality, coverage of CpG sites and known problematic genomic regions, DNA methylation quantification, detection of genomic alterations in GBM-associated genes, and finally applicability to haplotype reconstruction.

Results: 5-base sequencing provided flexible sequencing yield, highest basecalling quality and CpG coverage, supporting scalable and accurate genomic and epigenomic profiling, whereas ONT sequencing allowed for uniform coverage of problematic genomic regions and variants phasing. Both sequencing platforms showed high concordance with EPICv2 methylation measurements, with Pearson correlation coefficient 95% CI of 0.95-0.96 for 5-base and 0.94-0.95 for ONT. Measurement discrepancies were greater at intermediately methylated CpG sites and generally decreased with increasing sequencing depth. In genomic analyses, 5-base sequencing, due to substantially higher yield-to-cost ratio, outperformed ONT in somatic small variant detection. At the same time, ONT long reads enabled uniform sequencing of problematic genomic regions as well as effective haplotype reconstruction and detection of allele-specific *MGMT* promoter methylation. Finally, integration of 5-base and ONT data further improved phasing performance.

Conclusions: The evaluated sequencing platforms provide complementary capabilities for genomic and epigenomic profiling. 5-base offers sequencing scalability, high basecalling accuracy, and broad methylome coverage, whereas ONT improves interrogation of problematic genomic regions and enables haplotype-resolved methylation analysis. EPICv2 microarrays provide cost-effective, scalable and highly consistent methylation measurements but remain restricted to predefined CpG loci.

Keywords: 5-base, EPICv2, Oxford Nanopore Technologies, MGMT, DNA methylation

Background

Recent advances in short-read sequencing (SRS) and microarray technologies have enabled high-throughput molecular profiling across various levels of cell biology including: genome, epigenome, transcriptome and recently proteome (1,2). These technologies facilitate not only exploration of cancer biology but also open unprecedented opportunities to improve clinical cancer management (3,4). With wide availability of molecular profiling platforms, the detection of genomic alterations underlying cancer pathology has already become central to precision oncology, supporting cancer diagnosis, prognosis, disease monitoring, and therapeutic decision-making (5,6). In parallel to progress in genetic biomarkers research, the applicability of cancer-related DNA methylation alterations as biomarkers is being increasingly demonstrated. A prominent application of DNA methylation profiling is the classification of central nervous system (CNS) tumors, enabling more precise molecular subtyping and improved diagnostic resolution beyond conventional histopathological assessment (7,8). Also, our recent work demonstrated remarkable accuracy of machine-learning models based on DNA methylation biomarkers for primary tumors and metastases classification (9).

SRS and microarray-based technologies dominate genetic and epigenetic profiling in both research and diagnostics however, they are facing important technical limitations. In particular, SRS has limited utility for haplotype phasing. It also shows reduced accuracy for variant detection in problematic genomic regions (PGRs), especially those characterized by low sequence complexity or high sequence homology (10,11). Finally, the detection of short tandem repeats (STRs) and structural variants (SVs) using SRS remains challenging (12,13).

In epigenomic applications, specifically DNA methylation assessment, SRS-based approaches typically rely on bisulfite conversion, which 1) dilutes nucleotide diversity forcing increased PhiX spike-in (14), and 2) fragments DNA (15), both effects reduce sequence complexity leading to lowered mapping efficiency and

in consequence increased costs (16,17). At the same time, microarrays, while accurate and cost-effective, are limited to interrogating only a predefined subset of CpG sites in the genome (18,19).

The limitations of SRS have recently been at least partly addressed by improvements in library preparation chemistry and data processing algorithms. The incorporation of unique molecular identifiers (UMI) into libraries enable the grouping of reads derived from individual DNA molecules and generation of consensus sequences, thereby reducing PCR- and sequencing-derived errors (20). Moreover, recent progress in the field of pangenome references (21), as well as proximity-mapped reads technology improves mapping accuracy of short-reads in challenging genomic regions (22).

In DNA methylation research, the recent development of enzymatic, conversion-based 5-base sequencing has provided an alternative to standard SRS approaches, including whole-genome bisulfite sequencing (WGBS) and enzymatic methyl-seq (EM-seq). The key advantage of 5-base sequencing over WGBS and EM-seq is the preservation of library diversity, including both nucleotide diversity and insert integrity. By avoiding bisulfite-induced DNA degradation and retaining predominantly unmethylated cytosines in the human genome, 5-base sequencing enables more efficient and uniform genome-wide methylation profiling.

Importantly, the limitations of SRS and microarrays have, at least to some extent, been overcome by long-read sequencing (LRS) technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Especially ONT, after the introduction of a new generation of R10 nanopores and v14 chemistry, has been increasingly adopted in molecular studies due to superior accuracy in comparison to R9 nanopores (23), simplified workflows (24), real-time basecalling (25), the ability to detect various covalent DNA/RNA modifications (26), and growing bioinformatics ecosystem, centered around EPI2ME repository. Importantly, both PacBio and ONT enable sequencing of native DNA molecules, thereby reducing sample preparation related biases. Furthermore, sequencing of long DNA fragments significantly improves haplotypes assembly, which is particularly important for interpretation of the molecular

consequences of genomic alterations as well as for assessment of allele-specific DNA methylation. Nevertheless, LRS platforms remain limited by low sequencing yield, higher costs, and lower basecalling accuracy compared with SRS platforms (27).

Recent comparisons of short- and long-reads technologies for genetic variant assessment have shown that while short-read platforms remain highly accurate for small variant detection in vast majority of genomic regions, long-read technologies provide superior efficiency for variant calling in PGRs as well as for structural variant detection (28). Similar assessment of newest 5-base protocol based on short-reads, and long-reads technology for parallel genomic and epigenomic profiling of tumor samples has not been performed.

Therefore, in this study, we systematically evaluated SRS and LRS technologies for integrated genomic and epigenomic profiling of tumor DNA. We then benchmarked 5mC estimates generated by ONT and 5-base sequencing against EPICv2 microarrays, recognized as gold standard methylation screening platform (18,29). Importantly, technological benchmarking studies are often performed using deeply characterized reference cell lines (30,31). Although this strategy is well suited for assessing analytical parameters such as variant-calling sensitivity and specificity, it does not reflect the impact of sample heterogeneity as well as variability in input DNA quality and quantity.

Therefore, to provide a clinically relevant assessment of SRS and LRS performance, we decided to focus our study on 16 brain tumor samples initially diagnosed as glioblastoma (GBM), a highly heterogeneous tumor type characterized by both genomic and epigenomic alterations (32,33).

Results

EPICv2 microarrays provide highly accurate DNA methylation profiles

Illumina bead-based microarray technology has been widely adopted in epigenetic studies, largely because of its cost-efficiency, scalability (up to 3 024 samples per week on one instrument, according to the

manufacturer), reproducibility, and precision comparable to WGBS performed at approximately 100× coverage (18,29). This technology has been recognized as a robust source of DNA methylation profiling data not only for basic research, but also for the training and validation of machine-learning models, which have already demonstrated high accuracy in cancer detection and classification (7,9,34).

For these reasons, we used EPICv2 microarrays as the reference technology for head-to-head comparison with the 5-base and ONT sequencing platforms. Nonetheless, microarrays also suffer from technical limitations, including susceptibility to technical noise arising from minor differences in sample processing as well as slide and reagent lots, collectively referred to as batch effect. Therefore, to ensure the quality of DNA methylation data that we benchmarked both 5-base and ONT technologies against, we first assessed parameters characterizing microarray data quality including: fraction of probes passing filtering, signal intensity metrics as well as beta-values distribution for each of the samples processed in our study. Overall, all 16 microarrays (2 slides, eight microarrays each) used in this experiment provided methylation level measurements at uniform number of CpG sites, exceeding 94% of all CpG sites targeted by EPICv2 (Fig. 1A). Distributions of red-to-green channel intensity ratios (Fig. 1B) as well as distortion metric for the informative probes (Fig. 1C) were comparable (*p-value* > 0.05, Mann-Whitney-U test) between the two slides used in the experiment, indicating balanced signal acquisition across dye channels. These metrics confirm uniform signal properties and no detectable slide- or microarray-specific technical bias. To further evaluate the quality of methylation levels measurements generated by microarrays, we assessed the association between probes intensities transformed into methylation levels (β-values) and log2-transformed probes intensity (TI). This analysis showed consistent β-values to TI ratio across, both slides used in our experiment (Fig. 1D-E) and no visible slide-specific artifacts. The β-value distributions across all samples in the experiment displayed a characteristic bimodal pattern, with peaks corresponding to low -, and high-methylation levels (Fig. 1F). We also used multidimensional scaling (MDS) data decomposition method to assess global similarity of generated 16 DNA methylation profiles. The projection of β-values

from each sample in two-dimensional space using MDS and standardized Euclidean distance metric showed uniform distribution of methylation profiles, with no evidence of clustering suggesting the presence of slide-associated batch effect (Fig. 1G).

Finally, we used our recently published OPP platform for machine-learning–based cancer classification (9) to confirm the initial GBM diagnoses and determine the molecular subtypes of the tumors included in our study. The CNS tumor classifier v1.0 classified 14 of 16 samples as GBMs, one as a low-grade ganglioglioma, and one as a pituitary adenoma (Fig. 1H). To investigate the two discordant classifications, we reviewed the corresponding sample collection records and found that the specimen classified as a pituitary adenoma had been incorrectly labeled as GBM during sample collection. For the second discordant sample, we could not identify any procedural or labeling issue. Importantly, these two discordantly classified samples were excluded from subsequent GBM-specific analyses.

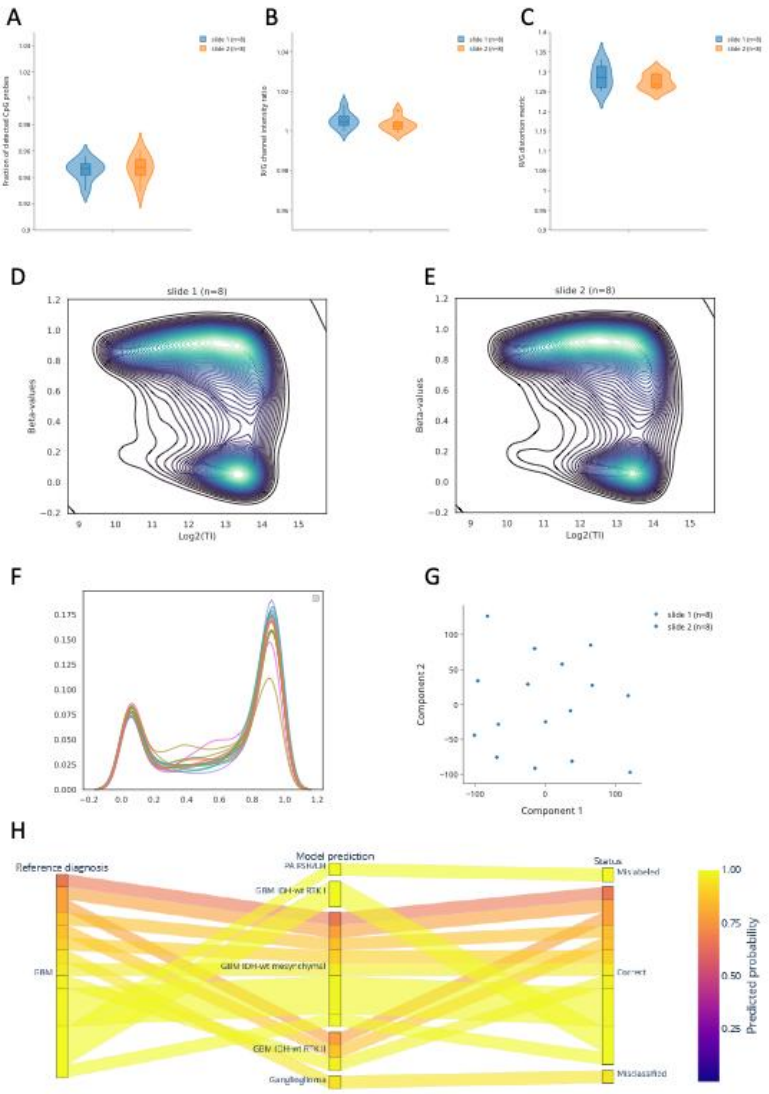


Figure 1. Results of quality assessment of DNA methylation data generated using two EPICv2 slides. **Fig. 1A-C:** Violin plot describing distribution of the fraction of detected probes (panel A), red-to-green channel intensity ratios (panel B), red-to-green channel intensity distortion metric (panel C) across samples stratified by slide. **Fig. 1D-E:** Two-dimensional density plot describing average association between β-values and log2-transformed probes intensity for slide 1 (panel D) and slide 2 (panel E). **Fig. 1F:** Kernel density plot describing distribution of β-values across samples. **Fig. 1G:** Scatter plot describing multi-dimensional scaling analysis of methylation profiles projected onto the first two components. **Fig. 1H:**

Sankey plot showing flow between reference diagnosis, model prediction, and final classification status, with color scale representing prediction probability.

Head-to-head comparison of DNA methylation profiles generated using 5-base, ONT and EPICv2 microarrays

Sequencing platforms differ in terms of 5mC quantification precision (35,36), which can mainly be attributed to varying sequencing depth (37), coverage uniformity (35), reaction efficacy under specific experimental conditions (38) for conversion-based assays including 5-base solution, or in case of ONT basecalling model accuracy (39).

Thus, we began assessment of quality of 5mC quantification between 16 samples sequenced with 5-base and 10 samples sequenced with ONT technologies from comparison of general sequencing performance metrics including sequencing yield expressed as the number of sequenced base pairs, basecalling quality measured by Phred scores and mapping efficacy in problematic genomic regions (PGRs) defined by The Genome in a Bottle consortium (40). As expected, 5-base technology in combination with 10B 300C flowcell, provided median > 200 Gbp yield per sample with average sequencing depth at the level of 56 (95% CI: 53-59), while ONT with PromethION P2 flowcells gave median yield of > 50 Gbp per sample and average depth of 23 (95% CI: 17-28) (Fig. 2A). The analysis of basecalling quality, using Phred scores (Q-score) representing the negative log-transformed probability of basecalling error, revealed substantial differences between the two technologies. With 5-base sequencing data consistently achieving high Q-score approaching Q40 (95% CI: 39.13–39.16) across all sequenced samples, and ONT showing lower, and more variable Q-score, with a mean value of Q19.3 (95% CI: 18.86–19.77) (Fig. 2B).

Next, we assessed read-mapping efficiency in known PGRs, including regions characterized by low sequence complexity, high homology, or duplicated segments. In this analysis we compared sequencing depth estimated globally at genome-wide level and for specific PGRs, next we defined mapping efficacy as

a ratio of PGRs sequencing depth to average genome-wide sequencing depth, assuming that values close to 1 reflect lack of systematic bias in reads mapping effectiveness for PGRs. The analysis confirmed that ONT provided more uniform coverage of problematic regions, with depth ratios close to 1. At the same time, 5-base sequencing showed reduced mapping efficacy in PGRs, which is consistent with known limitations of SRS (Fig. 2C).

To benchmark the sequencing-based approaches against the microarray, we first restricted the analysis to CpG sites uniformly assessed across all available samples for each platform, comprising 16 samples for 5-base and EPICv2 as well as 10 samples for ONT. Overall, 5-base, ONT, and EPICv2 enabled uniform methylation quantification at approximately 28 million, 8 million, and 788 138 CpG sites, respectively (Fig. 2D).

Subsequently, we focused on the intersection of 189 430 CpG sites uniformly quantified by all three platforms and compared the estimated β-value distributions. This analysis showed that all three technologies generated methylation profiles with the expected bimodal distribution, reflecting the predominance of hypo- and hypermethylated CpG sites in the human genome (Fig. 2E). Interestingly, ONT data showed a slightly broader β-value distribution, which may be related to the direct detection of DNA methylation without bisulfite or enzymatic conversion.

We also compared average DNA methylation levels across the analyzed subset of CpG sites between each of the technologies using multi-dimensional scaling (MDS) analysis. Interestingly, MDS revealed clustering driven by both sample and the type of technology (Fig. 2F). These findings suggest that platform-specific measurement and preprocessing effects contribute to the observed clustering, while the overall methylation signal remains broadly concordant. Notably, EPICv2-derived profiles showed the lowest variability of estimated methylation levels, whereas sample-to-sample variability increased in case of 5-base-, and ONT-derived methylation profiles (Fig. 2F).

Finally, we assessed concordance between 5-base and EPICv2 methylation profiles as well as ONT and EPICv2 profiles within previously described set of common CpG sites. This analysis was based on the linear association between data generated by the compared technologies, expressed as the Pearson correlation coefficient (r), and on Bland–Altman plots describing the association between mean methylation levels and differences in methylation levels measured by each of technologies. Across 10 matched samples and 189 430 common CpG sites, DNA methylation profiles generated by 5-base and ONT sequencing were highly concordant with EPICv2 microarray data, achieving Pearson correlation 95% CI: 0.95-0.96 for 5-base and 95% CI: 0.94-0.95 for ONT as well as absolute difference 95% CI: 0.075-0.083 for 5-base and 95% CI: 0.087-0.093 for ONT (per sample, pairwise-oriented analyses based on all CpG sites available for specific comparison are available in Supplementary Figures 1-16). Despite the overall concordance between platforms, these results are likely influenced by the abundance of hypo- and hypermethylated CpG sites in the human genome. Moreover, this analysis does not account for sequencing depth, which may affect measurement quality and consequently inter-platform concordance.

Therefore, we aimed to assess whether differences in β-values measured with 5-base and EPICv2, as well as with ONT and EPICv2, are uniform across the entire β-value range, and if any potential differences in β-values measured between platforms are sequencing depth-dependent. This analysis clearly showed that, for both 5-base (Fig. 3A) and ONT (Fig. 3B), the absolute differences between measurements generated by sequencing platforms and EPICv2 were relatively low for the most frequent hypo- and hypermethylated CpG sites, but increased in intermediately methylated regions. At the same time, analysis of the association between the absolute differences in methylation measurements obtained with EPICv2 and 5-base and sequencing depth showed that observed differences progressively decreased with increasing depth (Fig. 3C). A similar trend was observed for ONT, with differences decreasing up to approximately x30 coverage. However, beyond this point, the trend was no longer apparent, most likely due to the decreasing number of CpG sites passing the corresponding depth thresholds (Fig. 3D).

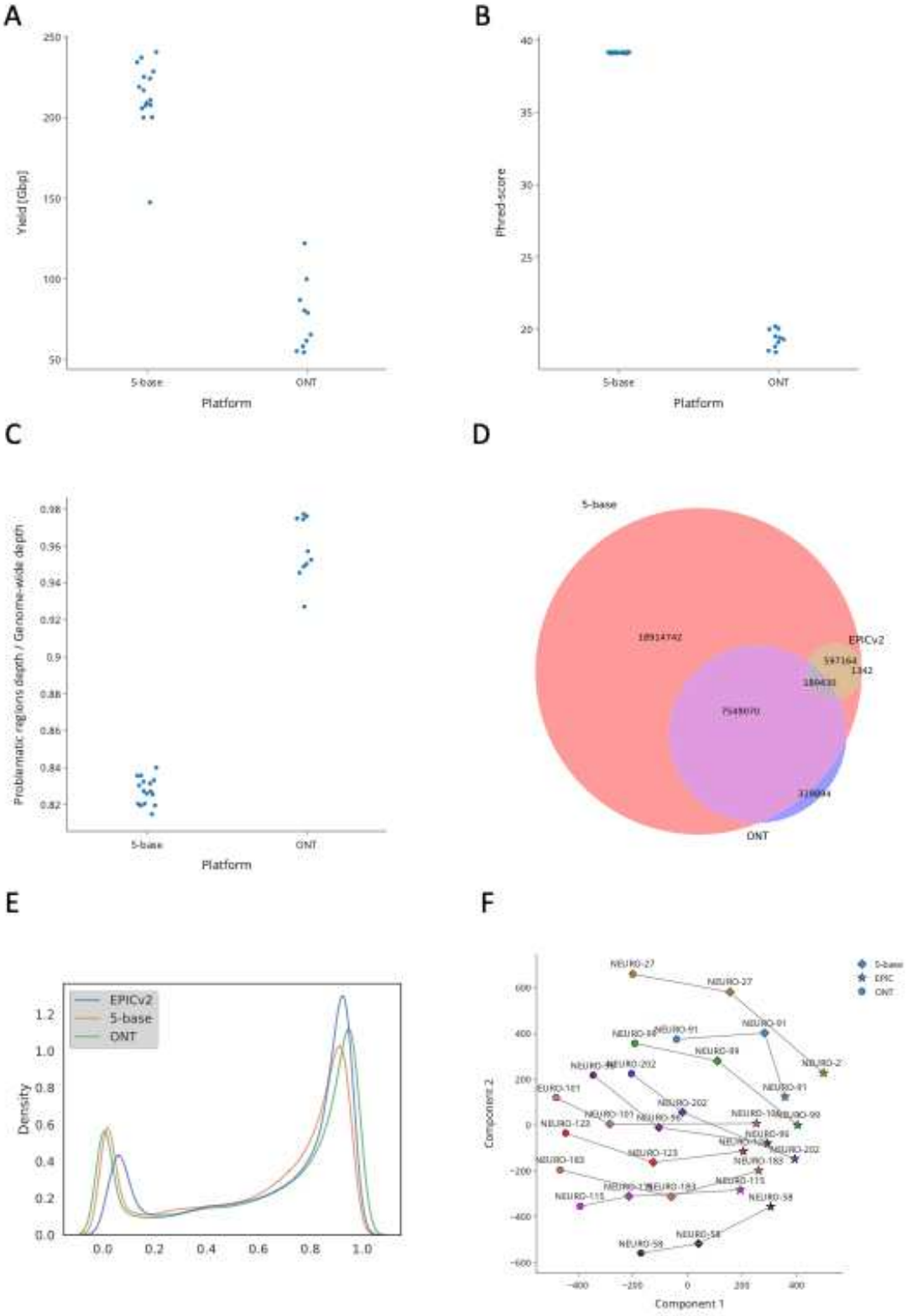

A
Yield [Gbp]
250
200
150
100
50
5-base
ONT
Platform
B
Phred-score
40
35
30
25
20
5-base
ONT
Platform
C
Problematic regions depth / Genome-wide depth
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
5-base
ONT
Platform
D
5-base
EPICv2
ONT
E
Density
EPICv2
5-base
ONT
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
F
Component 2
Component 1
5-base
EPIC
ONT

Figure 2: Assessment of concordance of DNA methylation profiles for 189 430 CpG sites common between EPICv2, 5-base and ONT platforms**. Fig. 2A:** Scatter plot describing distribution of sequencing yield (Gb) across samples stratified by sequencing platform. **Fig. 2B:** Scatter plot describing distribution of Phred scores across samples stratified by sequencing platform. **Fig. 2C:** Scatter plot showing ratio of problematic genomic regions depth to genome-wide depth stratified by sequencing platform. **Fig. 2D:** Venn diagram describing overlap of CpG sites quantified across 5-base, ONT and EPICv2 platforms, based on their genomic coordinates. **Fig. 2E:** Kernel-density estimation plot based on average methylation profiles generated with different platforms, showing expected bimodal distributions with peaks corresponding to hypo- and hypermethylated states. **Fig. 2F:** Scatter plot displaying similarity analysis, based on multi-dimensional scaling algorithm, of methylation profiles across distinct technological platforms. Each point represents single methylation profile, colors represent the origin of the sample and symbols platform used to generate data.

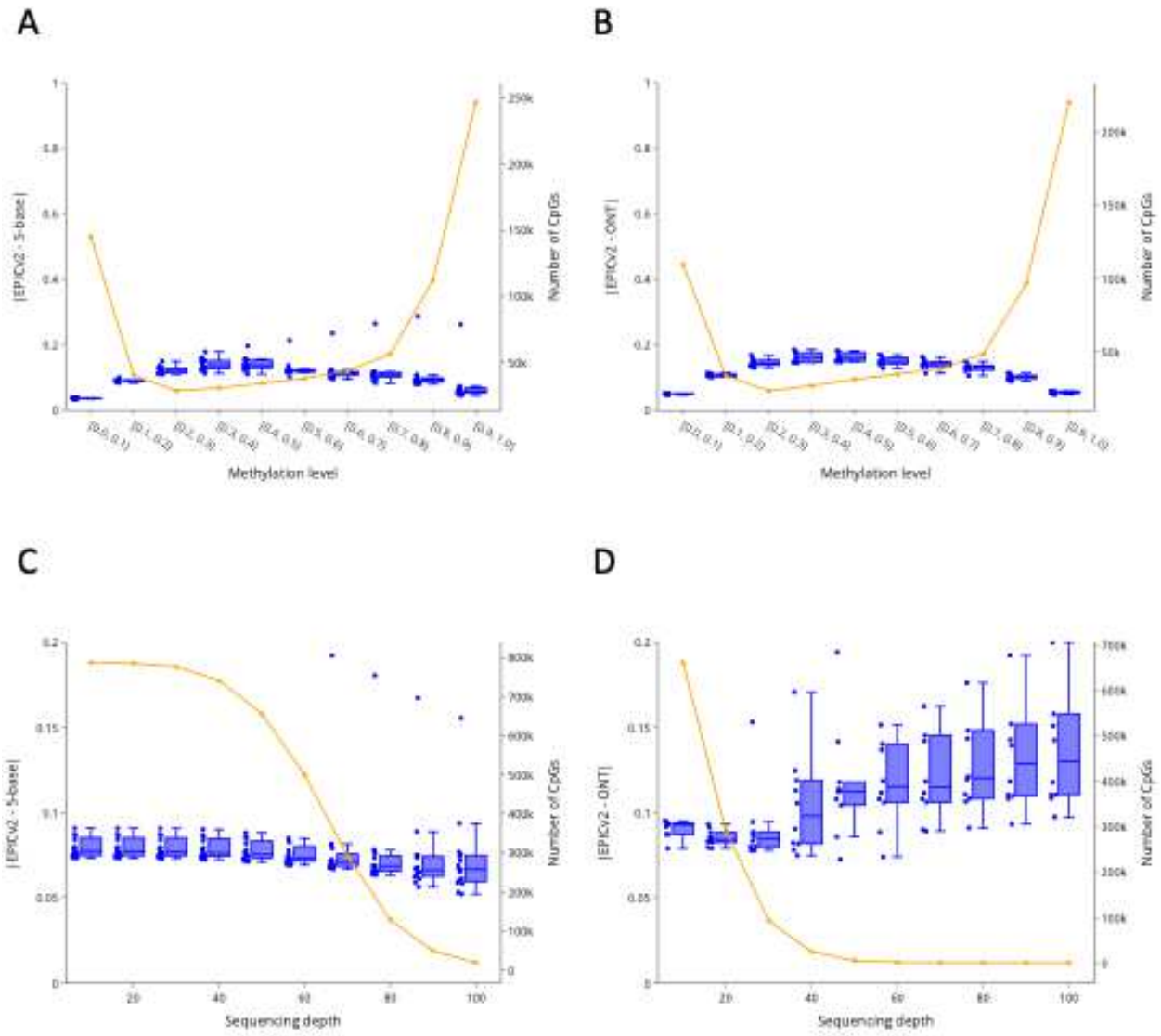


Figure 3. Methylation-level- and sequencing-depth-dependent differences between EPICv2 and sequencing-based DNA methylation measurements. **Fig. 3A-B:** Association between the absolute differences in β-values between EPICv2 and 5-base (panel A) as well as EPICv2 and ONT (panel B) across the binned β-value range. Box plots show the distribution of absolute differences for specific β-value intervals (bins), while the orange line represents the number of CpG sites in each interval, shown on the secondary y-axis. **Fig. 3C-D:** Association between sequencing depth and absolute differences in β-values between EPICv2 and 5-base (panel C) as well as EPICv2 and ONT (panel D). Box plots show the distribution

of absolute differences as a function of sequencing-depth, while the orange line represents the number of CpG passing depth threshold, shown on the secondary y-axis.

Head-to-head comparison of genomic profiles generated using ONT and 5-base sequencing

In parallel to epigenome assessment, we also aimed to compare the performance of 5-base and ONT sequencing for genomic profiling. However, a direct comparison of variant-calling accuracy between the two platforms was not possible because the true genetic background of the analyzed tumors was unknown. In addition, insufficient sequencing depth limited the sensitivity and reliability of somatic variant calling. Moreover, the absence of matched germline profiles, although representative of a realistic clinical setting, further complicated the interpretation and comparison of somatic variants.

Despite these limitations, we aimed to characterize the genomic profiles generated by both platforms and assess their consistency with the expected glioblastoma-specific mutational landscape. Firstly, we compared the overall characteristics of the somatic variant profiles called within coding regions generated by the two technologies. We observed that ONT yielded over 2x more somatic variants per sample than 5-base sequencing, with a mean of 183 variants (95% CI: 157-209) and 85 variants (95% CI: 75-95) for ONT and 5-base respectively. However, subsequent analysis of called variants, based on the Genome Aggregation Database (gnomAD) revealed that variants classified as somatic by the 5-base workflow were predominantly rare in population databases. In contrast, the set of variants identified with ONT and classified as somatic contained a substantial fraction of relatively common, therefore potentially misclassified germline variants (Fig. 4A).

After restricting the analysis to variants with gnomAD allele frequency below 0.001, we observed further differences in the composition of the putative somatic profiles. Although missense substitutions were the most frequent variant class for both technologies, somatic profiles obtained with 5-base sequencing were substantially more diverse, with non-missense variant classes collectively accounting for approximately

33-62% of detected variants across individual samples (Fig. 4B). In parallel, ONT-derived somatic profiles were dominated by missense substitutions, with non-missense variants comprising only approximately 6-17% of detected variants (Fig. 4C). The platforms also differed substantially in the distribution of variant allele frequencies (VAF). Across matched samples, somatic variants called by ONT had a median VAF of approximately 0.5, whereas the corresponding median for 5-base sequencing was approximately 0.2 (Fig. 4D). Consistent with this observation, MATH scores, reflecting the dispersion of VAFs within individual tumors and commonly used as a proxy for intratumoral heterogeneity (41,42), were significantly lower for ONT than for 5-base sequencing, with mean values of 27.9 and 39.6, respectively (Fig. 4E, *p-value* ≤ 0.05, Wilcoxon test). Notably, the mean MATH score calculated based on VAF obtained with 5-base sequencing was consistent with values previously reported for GBM based on whole-exome sequencing (46.86 ± 16.13) (42).

We next evaluated concordance between the two platforms using the Jaccard similarity index (JSI). Overall, JSI between putative somatic variant sets identified by each of the technologies was low, with JSI values ranging from 0 to 0.14. In contrast, when germline variants were analyzed using the same approach, concordance increased substantially, with JSI values ranging from 0.41 to 0.69 across assessed samples. Importantly, JSI values for both germline and somatic variant sets increased with increasing ONT sequencing depth (Fig. 4F), whereas a similar trend was not detectable in case of 5-base data (Fig. 4G). This association suggests that the lower ONT sequencing depth contributed substantially to the observed inter-platform discordance.

To determine whether observed differences between platforms affected detection of potentially clinically relevant alterations, we subsequently focused our analysis on 43 genes associated with GBM (43) and analyzed nine samples with a confirmed GBM diagnosis that had been sequenced using both technologies. ONT identified 15 somatic variants distributed across six out of nine samples, whereas 5-base sequencing identified 10 variants, also distributed across six out of nine samples (Fig. 4H). Only six variants were

concordantly classified as somatic by both technologies: *EGFR*.1934C>G (p.Ser645Cys) in sample NEURO-115, *PTPN11*.179G>T (p.Gly60Val) and *TP53*.530C>G (p.Pro177Arg) in NEURO-101, *PIK3CA*.1097C>G (p.Pro366Arg) in NEURO-99, *PTEN*.259C>T (p.Gln87Ter) in NEURO-96, and *PIK3CA*.1624G>A (p.Glu542Lys) in NEURO-123. The four variants detected exclusively by 5-base had relatively low VAFs, ranging from 0.04 to 0.19, supporting insufficient ONT sequencing depth as a potential explanation for their absence from the ONT calls. In parallel, the nine ONT-only somatic variants had substantially higher VAFs, ranging from 0.21 to 0.96. Importantly, all of these variants were also detected by 5-base but classified as germline, therefore excluded from the final 5-base somatic calls.

Finally, we assessed GBM-associated copy-number variations (CNVs), including focal *EGFR* amplification, chromosome 7 gain and chromosome 10 loss (44), using sequencing data generated for 14 and 9 samples with a confirmed GBM diagnosis for 5-base and ONT data, respectively. These results were subsequently compared with CNV profiles inferred from EPICv2 methylation-array data using conumee 2.0 (45) within the OPP (9) workflow (links to classification reports including CNV profiles are available in Supplementary Table 1). Visual inspection of the CNV profiles (Supplementary Table 2, Supplementary Figures 17-30) showed that the three approaches were broadly consistent and enabled the detection of at least one GBM-associated CNV category in the majority of assessed samples (Table 1).

platform/CNV type	***EGFR* focal amplification**	**ch7 gain**	**chr10 loss**
5-base	5/14 (36%)	10/14 (71%)	11/14 (79%)
EPICv2	7/14 (50%)	12/14 (86%)	11/14 (79%)
ONT	4/9 (44%)	8/9 (89%)	6/9 (67%)

Table 1: Comparison of the frequency of selected GBM-associated CNVs detected using 5-base, EPICv2 and ONT platforms.

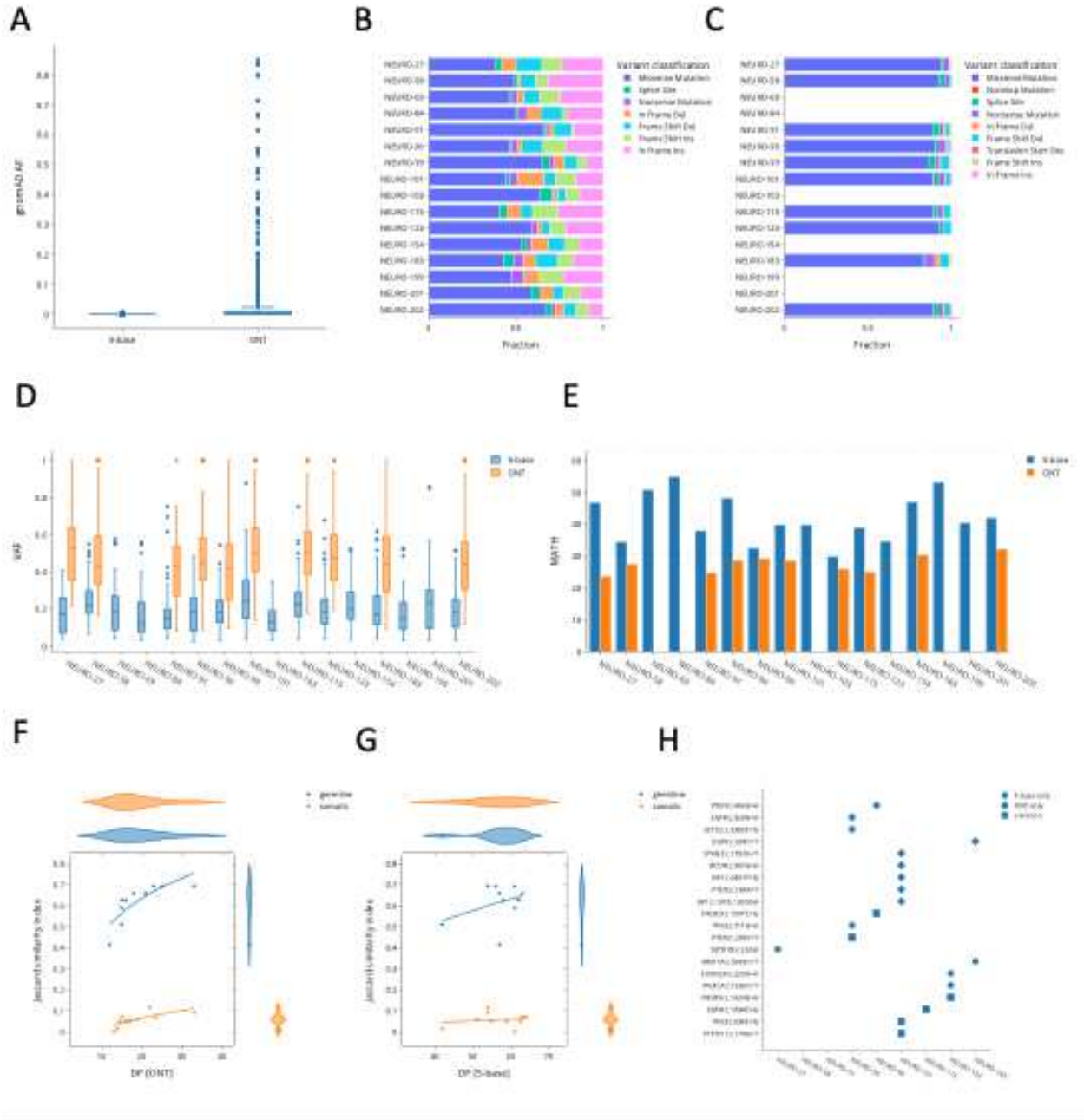


Figure 4: Comparison of genomic profiles generated using 5-base and ONT platforms. **Fig. 4A:** Distribution of gnomAD allele frequencies (gnomAD AF) for putative somatic variants identified by 5-base and ONT. **Fig. 4B-C:** Relative distribution of variant classes across samples for 5-base (panel B) and ONT (panel C). **Fig. 4D:** Distribution of variant allele frequencies (VAFs) for somatic variants detected by 5-base and ONT across assessed samples. **Fig. 4E:** Mutant-allele tumor heterogeneity (MATH) scores

calculated from somatic VAF distributions for each sample and sequencing platform. Dashed lines indicate the mean ± SD MATH score previously reported for GBM based on whole-exome sequencing. **Fig. 4F-G:** Association between sequencing depth and concordance of variant detection between platforms, quantified using the Jaccard similarity index, separately for germline and somatic variants. Concordance between platforms is shown as a function of median ONT sequencing depth (panel F) and median 5-base sequencing depth (panel G)**. Fig. 4H:** Somatic variants identified within 43 selected GBM-associated genes in nine samples with a confirmed GBM diagnosis sequenced using both platforms. Symbols indicate variants detected exclusively by 5-base sequencing, exclusively by ONT, or by both technologies.

Combination of short- and long-reads allows for effective haplotype-resolved DNA methylation profiling

Reads phasing, also referred to as haplotype assembly, is the process of reconstructing the allelic configuration of genetic variants at individual homologous chromosomes by assigning genomic alterations to their respective parental or even subclonal haplotypes. Phasing enables the analysis of variants inherited together, providing additional biological and clinically relevant context (46) as well as identification of allele-specific DNA methylation patterns (47).

Applicability of phasing based on short-read sequencing platforms is inherently limited, as haplotype reconstruction requires sequencing reads to span heterozygous variants, which is less likely for reads composed of up to 600 bp. In contrast, ONT sequencing can generate reads reaching even megabase-scale lengths, substantially increasing the probability of capturing multiple heterozygous loci within a single sequenced molecule. However, reliable haplotype assembly also requires highly accurate genotype calls, and as we have shown ONT-derived data are characterized by lower diversity of variant categories as well as overall basecalling accuracy compared with SRS. Therefore, following recommendations described in technical documentation of WhatsHap (48), we investigated whether combining 5-base genotype calls

with long ONT reads could improve haplotype reconstruction and enable more robust haplotype-resolved DNA methylation analysis.

An obvious limitation of this analysis was limited ONT ability for the assessment of somatic variants and therefore somatic haplotypes. Therefore, we restricted the analysis to biallelic heterozygous variants with allele fractions between 0.30 and 0.70, representing informative sites suitable for subsequent germline haplotypes assembly.

To assess the applicability of phasing based on combined 5-base and ONT data, we first performed haplotypes assembly solely based on ONT data, and heterozygous variants sequenced at a depth ≥ 10. We then repeated the analysis using long ONT-reads and variants identified from 5-base sequencing with a depth greater than 30. To compare these two phasing strategies, we calculated performance metrics for both approaches, including the number of identified haplotype blocks, median block length, maximum block length, and block contiguity expressed as the N50 metric (Fig. 5A-D). Overall, the number of blocks decreased significantly (*p-value* ≤ 0.05, Wilcoxon test) in case of hybrid approach, while remaining metrics did not change in terms of median values (*p-value* > 0.05, Wilcoxon test). However, the third-quartile and maximum values for median block length (Fig. 5B), maximum block length (Fig. 5C), and block N50 (Fig. 5D) were substantially higher with the hybrid approach. These results clearly show that ONT data, even at limited sequencing depth, is sufficient for effective haplotype assembly. Nevertheless, integration of long-, and short-reads can further improve overall efficacy of haplotype reconstruction.

Finally, we investigated whether assigning DNA methylation status to specific alleles can have potential clinical implications. We performed this analysis for *MGMT* promoter, a clinically relevant biomarker used to predict response to temozolomide (TMZ) chemotherapy in glioblastoma (49). Initial analysis of average methylation levels at *MGMT* promoter in 14 GBMs included in our study (Supplementary Figure 31) using 5-base data, indicated that in the majority of the samples *MGMT* promoter was either methylated (n=4) or not methylated (n=9). Interestingly, one sample (NEURO-99) displayed average methylation levels

indicative of partial (intermediate) methylation, which could reflect allele-specific 5mC patterns. The phasing based on combined 5-base and ONT strategy for this sample, clearly demonstrated that all reads assigned to one haplotype (HP1) were fully unmethylated, whereas only a subset of reads assigned to the second haplotype (HP2) carried methylated cytosines within the *MGMT* promoter region (Fig. 5E). These findings suggest that the unmethylated haplotype may retain the potential for *MGMT* expression, and consequently resistance of tumor cells to TMZ, although this needs to be confirmed with allele-specific expression and protein-level analyses**.**

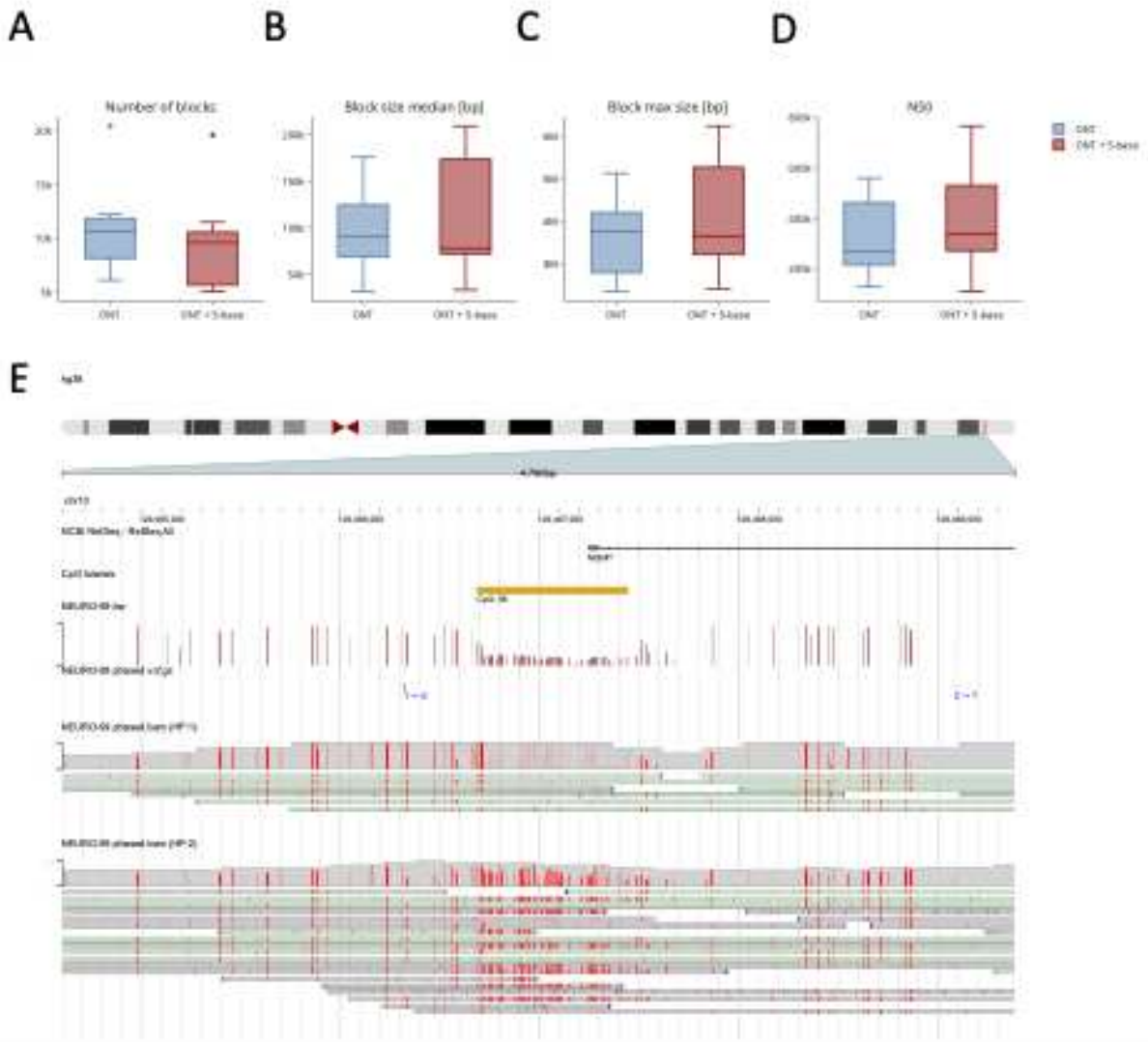


Figure 5. Integration of short- and long-read sequencing improves phasing efficacy and enables haplotype-oriented assessment of 5mC. **Fig. 5A-D:** Comparison of phasing efficacy metrics obtained using ONT and a hybrid (ONT + 5-base) approach, including the number of haplotype blocks (panel A), median block size (panel B), maximum block size (panel C), and block continuity expressed as N50 (panel D). **Fig. 5E:** Representative genomic region centered on the *MGMT* promoter for NEURO-99 sample. Tracks (in order from top to bottom) show genomic coordinates and gene annotation, CpG island position, global DNA methylation levels estimated from 5-base sequencing, phased genomic variants from 5-base, and ONT long reads assigned to haplotype 1 (HP1) and haplotype 2 (HP2). CpG methylation status in individual ONT reads is shown as blue for unmethylated CpG sites and red for methylated CpG sites.

Study limitations

Our study has several major limitations that should be considered when interpreting the results. First, ONT sequencing was performed for only 10 of the 16 samples, limiting the sample size and completeness of the cross-platform comparison for this technology. Second, only 2 of the 10 ONT flow cells (NEURO-91 and NEURO-202) generated sufficient data to achieve a mean genome-wide coverage exceeding x30. The lower sequencing depth obtained for the remaining samples reduced the precision and sensitivity of downstream analyses, especially in case of somatic variant detection, where adequate depth is critical for identifying low-VAF variants. Third, although EPICv2 microarrays were used as a reference for DNA methylation quantification, they interrogate only a predefined subset of CpG loci. Consequently, cross-platform methylation concordance was assessed only based on a fraction of CpG sites spanning human genome. Finally, our study did not include technical replicates and therefore does not address the technical reproducibility of the evaluated platforms. Such assessment of intra-platform reproducibility remains particularly important in the context of potential clinical implementation and compliance with CE-IVD requirements.

Discussion

Simultaneous assessment of the genome and epigenome within a single assay offers several benefits. First, it reduces experimental time and costs by enabling both molecular modalities to be characterized from the same sample, in exactly the same biological context, and within a unified analytical workflow. Second, combining genetic and epigenetic information substantially increases the dimensionality of molecular information available for downstream analyses, which may be particularly valuable not only for biological interpretation but also for machine-learning-based solutions.

Over the last years, substantial progress has been made in the development of high-throughput molecular technologies. Nevertheless, many platforms, particularly recently introduced, remain insufficiently evaluated in clinically relevant settings. Therefore, systematic evaluation of emerging omics platforms in clinically relevant settings is essential to determine their practical utility, identify technology-specific limitations, and define the applications for which a particular platform is best suited.

In this study, we systematically evaluated the new 5-base solution and compared it with ONT sequencing as well as EPICv2 microarrays using glioblastoma specimens, which constitute challenging and clinically relevant material characterized by substantial biological heterogeneity.

Our results demonstrated that both sequencing platforms have distinct strengths and limitations. Their suitability should therefore be considered in the context of the intended application, taking into account DNA quality and availability, required sequencing depth, genomic variant classes of interest, CpG coverage, and the need for long-range information.

Specifically, in our analyses, the 5-base platform provided flexible sequencing yield, higher base-calling accuracy, more uniform genome-wide CpG coverage, and the highest concordance of 5mC profiles with EPICv2 microarrays. Nevertheless, 5-base sequencing remains subject to the inherent limitations of short-read technologies, particularly, reduced mapping efficiency in complex genomic regions (50) and limited suitability for fully haplotype-resolved analyses (51).

At the same time, ONT provided more efficient mapping in challenging genomic regions. In addition, ONT-based 5mC quantification showed high concordance with results generated using EPICv2 microarrays. Importantly, from both biological and clinical perspectives, ONT long reads enable haplotype-oriented genomic and epigenomic analyses by providing long-range information that is limited in conventional short-read sequencing. However, our results, together with previous studies, indicate that the platform is characterized by lower basecalling quality scores, typically below Q30 (52), as well as limited precision and recall for somatic indels in tumor-only mode (53). Furthermore, under our one-flow-cell-per-sample design, the lower per-sample coverage achieved with ONT constrained the sensitive detection of low-VAF variants and makes high-depth tumor WGS less efficient (50).

In this study, as a reference for human methylome assessment, we used EPICv2 microarrays, which interrogate a predefined set of CpG loci spanning regulatory regions across the human genome. Our results confirmed that EPICv2 generates methylation profiles highly comparable to deep sequencing-based approaches (29), while remaining cost- and time-efficient particularly in large-scale studies.

Interestingly, based on previously described recommendations (48), we also confirmed that combining high-quality genotypes derived from 5-base sequencing with long-read ONT data provides an effective hybrid framework for integrated, haplotype-aware genomic and epigenomic analyses. In this approach, 5-base data provide reliable genotype calls and high-quality 5mC quantification, whereas ONT reads enable long-range phasing and haplotype-resolved 5mC profiling. Using this hybrid approach, we were able to identify an allele-specific *MGMT* promoter methylation pattern that would remain inaccessible using conventional bulk methylation analyses. These findings highlight the potential biological and clinical value of integrating short- and long-read sequencing technologies in future precision oncology workflows, although this approach requires further validation.

Overall, our study suggests that 5-base sequencing offers the most balanced performance for genome-wide integrated genomic and epigenomic profiling, while ONT adds complementary long-range information that is particularly valuable for challenging genomic regions and haplotype-resolved analyses.

Conclusions

Our findings demonstrate that 5-base and ONT sequencing offer complementary strengths for simultaneous genomic and epigenomic profiling. Both platforms showed high concordance with EPICv2 microarrays in DNA methylation quantification, although differences were observed at intermediately methylated CpG sites and were influenced by sequencing depth. Under the evaluated experimental conditions, 5-base sequencing provided substantially greater sequencing yield, broader CpG coverage, higher basecalling accuracy, and more reliable somatic small variants profiling. In parallel, ONT demonstrated more uniform coverage of problematic genomic regions and enabled haplotype-aware analyses. Importantly, combining 5-base-derived genotypes with ONT long reads improved selected phasing metrics and enabled haplotype-resolved assessment of *MGMT* promoter methylation. Together, these results support 5-base sequencing as a scalable approach for comprehensive molecular profiling, with ONT providing complementary capabilities for investigating complex genomic regions and allele-specific epigenetic alterations.

Materials and Methods

Cohort characteristics

The study cohort comprised 16 patients initially diagnosed as glioblastomas (after data interrogation two samples were reassigned, for details see: **EPICv2 microarrays provide highly accurate DNA methylation profiles** section), including 8 females (50%) and 8 males (50%). At the time of study, 11 patients (68.8%) were deceased and 5 (31.2%) were alive. Methylation status of the *MGMT* promoter was assessed using

a PCR-based assay with promoter methylation confirmed in 8 of 15 samples (53.3%). Necrosis was observed in 14 of 16 tumors, anaplastic features in all 14 cases for which this information was available, and endothelial proliferation in 8 evaluable cases. Polymorphic cells were reported in 11 tumors, while hypercellularity was documented in 3 cases. Ki-67 proliferation indices were available for 14 tumors and ranged from 10% to 90%. Detailed characteristics of the cohort are provided in Supplementary Table 1.

Sample collection procedure

Brain tumor specimens were collected during routine neurosurgical resections performed at the University Clinical Hospital No. 1, Szczecin, Poland, as a part of treatment. Tissue fragments intended for molecular analyses were frozen at −80°C within 2 hours from the surgery.

DNA extraction, quantification and quality control

A total of 25 mg of tissue collected from multiple regions of the tumor mass was used to ensure representative contribution of spatially distinct tumor areas to the extracted DNA. High molecular weight (HMW) DNA was isolated using the MagAttract® HMW DNA Kit (Qiagen, cat. no. 67563). DNA concentration was measured using the Quantifiler™ Trio DNA Quantification Kit (Thermo Fisher Scientific, cat. no. 4482910) on a QuantStudio™ 5 Real-Time PCR System (Thermo Fisher Scientific). The DNA purity was assessed using BioTek Synergy H1 microplate reader (BioTek Instruments) according to the standard recommendations by calculating the ratios of corrected absorbance at 260/280 nm and 260/230 nm. Additionally, DNA integrity was evaluated using the Genomic DNA ScreenTape Assay (Agilent Technologies, cat. no. 5067-5365, 5067-5366) on a 4150 TapeStation System (Agilent Technologies) and expressed as the DNA Integrity Number (DIN). The DNA degradation was also assessed using the Quantifiler™ Trio DNA Quantification Kit (Thermo Fisher Scientific, cat. no. 4482910), which provides a ratio of short-to-long DNA

fragments expressed as Degradation Index (DI). All collected parameters, describing DNA concentration, purity, and integrity are available in Supplementary Table 1.

EPICv2 data generation and preprocessing

All samples were processed directly following the manufacturer's recommendations. Briefly, 550 ng of genomic DNA dissolved in 20 µl was used for bisulfite conversion using the EZ-96 DNA Methylation-Lightning™ MagPrep Kit (Zymo Research, cat. no. D5046). The Infinium Methylation EPIC v2.0 Kit (Illumina, cat. no. 20087707) was used for bisulfite-treated DNA methylation profiling using an automated liquid handling system - Infinium RoMa platform (Tecan). Raw IDAT files were generated using the iScan™ System (Illumina). IDAT files were processed using the *sesame* (54) R package with settings recommended for EPICv2 and subsequently annotated using Illumina EPICv2 A2 v2 manifest. Probes previously identified by (19), that do not map to their target location as well as probes not uniformly available across all samples used in this experiment were removed from analysis. All metrics describing EPIC v2 data quality are available in Supplementary Table 1.

5-base data generation and preprocessing

All samples were processed directly following the manufacturer's recommendations. Briefly, 100 ng of genomic DNA in a volume of 55 µl was fragmented using a Covaris M220 Focused-ultrasonicator (Covaris) at a setpoint temperature of 20°C, with an average incident power of 10W, peak incident power of 50W, 20% duty factor, 200 cycles per burst, and 50 seconds of treatment time. After sonication, DNA fragment size distribution was assessed using the 4150 TapeStation System (Agilent Technologies) with the High Sensitivity D1000 ScreenTape Assay (Agilent Technologies, cat. no. 5067-5584, 5067-5585), and fragment size of approximately 450 bp was confirmed for all the DNA samples. Subsequently, library preparation was performed according to the manufacturer's instructions using the Illumina 5-base DNA Prep (Illumina,

cat. no. 20140364) and the dedicated protocol (Illumina, cat. no. 200073497 v00), with Illumina Unique Dual Indexes LT (Illumina, cat. no. 20083060). Library concentration and quality were assessed fluorometrically using Qubit™ Flex Fluorometers (Thermo Fisher Scientific) and Qubit™ dsDNA Broad Range (BR) Assay Kit (Thermo Fisher Scientific, cat. no. Q32850) and by automated electrophoresis on a 4150 TapeStation System (Agilent Technologies) using the D1000 ScreenTape Assay (Agilent Technologies, cat. no. 5067-5582, 5067-5583). Libraries were normalized to 2 nM and pooled to generate a library pool. Subsequently, libraries were diluted to 90 pM and sequenced on a MiSeq i100 Plus Platform (Illumina) using a 5M 300-cycle flow cell for quality control and re-pooling calculations. Finally, libraries were repooled according to the obtained results, re-quantified using the Qubit™ dsDNA High Sensitivity (HS) Assay Kit (Thermo Fisher Scientific, cat. no. Q32854) and sequenced on a NovaSeq X Plus platform (Illumina) using a 10B 300-cycle flow cell at a loading concentration of 130 pM and a 151×10×10×151 (U7N1Y143) cycles configuration. All subsequent data preprocessing steps, including alignment to multigenome hg38 reference, variant calling as well as 5mC quantification were performed using DRAGEN SOMATIC v4.4.6 within Illumina Connected Analytics platform. All metrics describing sequencing quality and efficacy are available in Supplementary Table 1.

Oxford Nanopore Sequencing data generation and preprocessing

Libraries were prepared using Ligation Sequencing Kit (Oxford Nanopore Sequencing, cat. no. SQK-LSK114) and protocol GDE_9161_v114_revAC_24Sep2025 with NEBNext Companion Module v2 for Oxford Nanopore Technologies Ligation Sequencing (New England Biolabs, cat. no. E7672S) according to the manufacturer's recommendation. 1000 ng of gDNA was used as an input for library preparation, after repeated quantification using Qubit dsDNA HS Assay Kit (Q32854, Thermo Fisher Scientific), as suggested by the manufacturer. Each library was sequenced on a single PromethION flow cell - FLO-PRO114M (Oxford

Nanopore Sequencing) for 72h using a maximum of 300 ng of the prepared library as an input for the flow cell (loaded DNA amount varied across libraries from 83 ng to 300 ng).

For the real-time quality control of the sequencing, initial basecalling was performed with the MinKNOW platform (Promethion 2 release), using the HAC (high accuracy) Dorado model version 5.2.0 with modified bases calling on 5mC cytosines in CG motifs. The HAC delivered data was also used to report preliminary findings and plan the further analysis.

The basecalling with the use of SUP (super accuracy) Dorado models, was guided by the limitations of the EPI2ME wf-somatic-variation v1.5.1 which allowed us to use only version 4.2.0 (with 5mC calling in CG motifs) and took place on the PromethION 2 Integrated, using its Tesla A100 GPU.

Basecalling employed the super-accurate model (dna_r10.4.1_e8.2_400bps_sup@v4.2.0). All subsequent data preprocessing steps, including alignment to hg38 reference, variant calling as well as 5mC quantification were performed using wf-somatic-variation v1.5.1 from EPI2ME repository. Finally, all metrics describing sequencing quality and efficacy are available in Supplementary Table 1.

Molecular diagnosis confirmation and estimation based on EPICv2 microarrays

We used our algorithms implemented in the OPP platform (9) to confirm histopathological diagnosis and estimate copy number variation (CNV) for each sample. Specifically, raw IDAT files were uploaded through the OPP web application (www.opp.pum.edu.pl), and the tumor class for each sample was predicted using the “CNS Tumors Classifier v1.0” model. CNV profiles were generated using conumee 2.0 (45) as a part of inference pipeline. Finally, GBM-specific CNVs, including chromosome 7 gain, focal *EGFR* amplification, and chromosome 10 loss were assessed by visual inspection of the CNV plots generated by the platform. Links to classification reports are available in Supplementary Table 1.

DNA methylation data processing

All datasets generated with sequencing platforms were converted to beta-value scale and represented using 1-based, unstranded genomic notation to ensure comparability with EPICv2 microarrays. All subsequent analyses were conducted using common libraries for data processing, analysis and visualization including polars, pandas, numpy, scipy and plotly within Python 3.11 environment.

Genomic data processing

VCF files generated using either the EPI2ME wf-somatic-variation v1.5.1 or the DRAGEN somatic v4.4.6 pipeline were filtered with bcftools (55) to retain variants that passed the workflow-specific quality filters and were supported by a minimum sequencing depth of 30× for 5-base or 10× for ONT data.

For 5-base sequencing, germline and somatic variants were distinguished based on the "GermlineStatus" INFO field. Variants annotated as "Somatic", "Somatic Putative CH", or "Somatic Hotspot" were classified as somatic, whereas all remaining variants were categorized as germline. Importantly, because the analysis was performed in tumor-only mode, this classification was inferred primarily from population-level reference databases, including gnomAD and the 1000 Genomes Project, rather than from a matched healthy sample. For ONT data, variant stratification was based on the "NonSomatic" FILTER status. Variants marked as "NonSomatic" were classified as germline, reflecting a non-somatic prediction derived from the reference resources incorporated into the EPI2ME workflow, including gnomAD, dbSNP, and the 1000 Genomes Project.

Prefiltered VCF files were subsequently annotated with the Variant Effect Predictor - VEP (56) against the GRCh38 reference genome using VEP cache version 115, restricted to canonical transcripts, and converted to Mutation Annotation Format (MAF). The MAF files were subsequently analyzed using the maftools R package (57). Copy-number variation profiles were generated based on data from DRAGEN somatic v4.4.6 for the 5-base sequencing data and SAVANA v1.3.8 (58) for ONT sequencing. For each sample, bin-level

log2-normalized copy-ratio values as well as segmented copy-ratio estimates were assessed. Analyses were restricted to chromosomes 1–22, X, and Y. Segment mean values, reported on a linear copy-ratio scale, were transformed to the $\log_2$ scale. Genomic bins were annotated using gene coordinates from the Ensembl GRCh38 release 115 GTF file. A bin was assigned to a gene when at least 25% of bin overlapped the gene coordinates. When multiple bins overlapped the same gene, the gene-level copy-ratio signal was summarized as the median log2-normalized ratio across all overlapping bins, thereby reducing the influence of individual extreme bin-level measurements. Finally, the status (gain or loss) of specific CNVs was assigned by visual inspection of the CNV plots.

Haplotype assembly

Haplotype assembly was performed using WhatsHap (48). Briefly, VCF files were filtered with bcftools (55) to retain autosomal, biallelic heterozygous SNV marked as “PASS” (in case of 5-base) or “NonSomatic” (in case of ONT), with a minimum sequencing depth of x30 (5-base) and x10 (ONT), with variant allele fraction between 0.30 and 0.70. The selected variants were then phased using the corresponding alignment files and reference genome, with “--merge-reads”, “--ignore-read-groups” and “--distrust-genotypes” flags. Finally, sequencing reads were assigned to the reconstructed “HP1” or “HP2” haplotypes.

List of abbreviations

5mC: 5-methylcytosine, AF: Allele frequency, BR: Broad range, CE-IVD: CE-marked in vitro diagnostic, CI: Confidence interval, CNS: Central nervous system, CNV: Copy-number variation, CpG: Cytosine–phosphate-guanine dinucleotide, DI: Degradation index, DIN: DNA integrity number, DNA: Deoxyribonucleic acid, DRAGEN: Dynamic Read Analysis for GENomics, EM-seq: Enzymatic methyl-seq, EPICv2: Infinium MethylationEPIC v2.0 BeadChip, GBM: Glioblastoma, GEO: Gene Expression Omnibus, gnomAD: Genome Aggregation Database, GPU: Graphics processing unit, HAC: High accuracy, HMW: High-

molecular-weight, HP1/HP2: Haplotype 1/Haplotype 2, HS: High sensitivity, IDAT: Illumina intensity data file, JSI: Jaccard similarity index, LRS: Long-read sequencing, MAF: Mutation Annotation Format, MATH: Mutant-allele tumor heterogeneity, MDS: Multidimensional scaling, ONT: Oxford Nanopore Technologies, PacBio: Pacific Biosciences, PGR: Problematic genomic region, Q-score: Phred quality score, SD: Standard deviation, SNV: Single-nucleotide variant, SRS: Short-read sequencing, STR: Short tandem repeat, SUP: Super accuracy, SV: Structural variant, TI: Log_2-transformed probe intensity, TMZ: Temozolomide, UMI: Unique molecular identifier, VAF: Variant allele frequency, VCF: Variant Call Format, VEP: Variant Effect Predictor, WGBS: Whole-genome bisulfite sequencing, WGS: Whole-genome sequencing.

Declarations

Ethics approval and consent to participate

The study was approved by Bioethical Committee of Pomeranian Medical University in Szczecin. All patients provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

DNA methylation profiles generated using the 5-base, ONT, and EPICv2 platforms were deposited in the Gene Expression Omnibus (GEO) database under accession numbers GSE347279 (5-base), GSE347388 (ONT), and GSE346968 (EPICv2).

Competing interest

T.K.W. has received honoraria from Illumina. Supporting entities had no influence on the study, the interpretation of the results, or the content of the manuscript, and did not review or approve the manuscript prior to submission.

Funding

This study was conducted by Regional Center for Digital Medicine at Pomeranian Medical University in Szczecin, agreement no. 2023/ABM/02/00006-00, funded by the Medical Research Agency.

Authors' contributions

J.B.: conceptualization, formal analysis, data curation, and writing. K.E.S: conceptualization, clinical data acquisition, data curation. M.S.: conceptualization, samples processing, laboratory experiments. K.S.: conceptualization, samples processing, laboratory experiments. K.M.: conceptualization, samples processing, laboratory experiments. M.K: data curation. P.P. data curation. T.v.d.W.: clinical sample collection, data curation. J.D.: samples processing, laboratory experiments. B.G. data curation, formal analysis. J.B.: clinical sample collection, clinical data acquisition. B.P.: clinical sample collection, clinical data acquisition. P.B.: clinical sample collection, clinical data acquisition. B.S.: clinical sample collection, clinical data acquisition. L.S.: clinical sample collection, clinical data acquisition, supervision. P.C.: conceptualization, supervision, funding acquisition. A.O.: conceptualization, supervision, funding acquisition, T.K.W.: conceptualization, supervision, funding acquisition, resources, and writing.

Acknowledgments

The authors gratefully acknowledge Illumina and Analityk Genetyka for their support and valuable assistance throughout the study.

References

1. Polonis K, Blommel JH, Hughes AEO, Spencer D, Thompson JA, Schroeder MC. Innovations in Short-Read Sequencing Technologies and Their Applications to Clinical Genomics. Clin Chem. 2025 Jan 3;71(1):97–108. doi:10.1093/clinchem/hvae173 PubMed PMID: 39749506.

2. De Souza BJ, Wolf MS, Leonard JR, Yu J, Pineda JA. Aptamer-based proteomics in pediatric patients with severe traumatic brain injury: a pilot study. Pediatr Res. 2026 Feb 19. doi:10.1038/s41390-026-04798-8

3. Matrana M, Shukla V, Kingsbury D, Poliak M, Lipton J, McMillin M, et al. Real-world data and clinical experience from over 100,000 multi-cancer early detection tests. Nat Commun. 2025 Oct 31;16(1):9625. doi:10.1038/s41467-025-64094-7 PubMed PMID: 41173830; PubMed Central PMCID: PMC12578847.

4. Najafi N, Sasaninia K, Najafi Z, Babakhan Vartanian C, Babakhan Vartanian K, Dungore F, et al. Predictive Genetic Testing in the Cancer Management and Prevention. J Cancer Sci Clin Ther. 2025;9(4):156–67. doi:10.26502/jcsct.5079274 PubMed PMID: 41541272; PubMed Central PMCID: PMC12802875.

5. Zhou Y, Tao L, Qiu J, Xu J, Yang X, Zhang Y, et al. Tumor biomarkers for diagnosis, prognosis and targeted therapy. Signal Transduct Target Ther. 2024 May 20;9(1):132. doi:10.1038/s41392-024-01823-2 PubMed PMID: 38763973; PubMed Central PMCID: PMC11102923.

6. Gu M, Zhang WZ, Fitzgerald RC, Frankell AM, Vassiliou GS. Harnessing genomics for early cancer detection, risk stratification and prevention. Nat Genet. 2026 Apr;58(4):704–16. doi:10.1038/s41588-026-02505-1 PubMed PMID: 41772265.

7. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018 Mar 22;555(7697):469–74. doi:10.1038/nature26000 PubMed PMID: 29539639; PubMed Central PMCID: PMC6093218.

8. Sill M, Schrimpf D, Patel A, Sturm D, Jäger N, Sievers P, et al. Advancing CNS tumor diagnostics with expanded DNA methylation-based classification. Cancer Cell. 2026 Feb 9;44(2):340-354.e2. doi:10.1016/j.ccell.2025.11.002 PubMed PMID: 41349541.

9. Bińkowski J, Wojdacz TK. DNA methylation biomarkers-based pan-cancer classifier: predictive modeling for cancer classification. Genome Med. 2026 May 19;18(1):66. doi:10.1186/s13073-026-01650-w PubMed PMID: 42152108; PubMed Central PMCID: PMC13185202.

10. Conlin LK, Aref-Eshghi E, McEldrew DA, Luo M, Rajagopalan R. Long-read sequencing for molecular diagnostics in constitutional genetic disorders. Hum Mutat. 2022 Nov;43(11):1531–44. doi:10.1002/humu.24465 PubMed PMID: 36086952; PubMed Central PMCID: PMC9561063.

11. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. Front Genet. 2019;10:426. doi:10.3389/fgene.2019.00426 PubMed PMID: 31134132; PubMed Central PMCID: PMC6514244.

12. Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. Hum Genome Var. 2024 Apr 17;11(1):18. doi:10.1038/s41439-024-00276-x PubMed PMID: 38632226; PubMed Central PMCID: PMC11024196.

13. Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. Sequencing and characterizing short tandem repeats in the human genome. Nat Rev Genet. 2024 Jul;25(7):460–75. doi:10.1038/s41576-024-00692-3 PubMed PMID: 38366034.

14. Nair SS, Luu PL, Qu W, Maddugoda M, Huschtscha L, Reddel R, et al. Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten. Epigenetics Chromatin. 2018 May 28;11(1):24. doi:10.1186/s13072-018-0194-0 PubMed PMID: 29807544; PubMed Central PMCID: PMC5971424.

15. Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. PloS One. 2018;13(6):e0199091. doi:10.1371/journal.pone.0199091 PubMed PMID: 29902267; PubMed Central PMCID: PMC6002050.

16. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 2018 Mar 15;19(1):33. doi:10.1186/s13059-018-1408-2 PubMed PMID: 29544553; PubMed Central PMCID: PMC5856372.

17. Otto C, Stadler PF, Hoffmann S. Fast and sensitive mapping of bisulfite-treated sequencing data. Bioinformatics. 2012 Jul 1;28(13):1698–704. doi:10.1093/bioinformatics/bts254 PubMed PMID: 22581174.

18. Kaur D, Lee SM, Goldberg D, Spix NJ, Hinoue T, Li HT, et al. Comprehensive Evaluation of The Infinium Human MethylationEPIC v2 BeadChip. Epigenetics Commun. 2023;3(1):6. doi:10.1186/s43682-023-00021-5 PubMed PMID: 38455390; PubMed Central PMCID: PMC10919401.

19. Peters TJ, Meyer B, Ryan L, Achinger-Kawecka J, Song J, Campbell EM, et al. Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling. BMC Genomics. 2024 Mar 6;25(1):251. doi:10.1186/s12864-024-10027-5 PubMed PMID: 38448820; PubMed Central PMCID: PMC10916044.

20. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017 Mar;27(3):491–9. doi:10.1101/gr.209601.116 PubMed PMID: 28100584; PubMed Central PMCID: PMC5340976.

21. Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. Nat Biotechnol. 2025 Jul;43(7):1177–91. doi:10.1038/s41587-024-02382-1 PubMed PMID: 39455800; PubMed Central PMCID: PMC12022141.

22. Bertoli L, Orlandi M, Timis AI, Lucchini F, Carta G, Limongi AR, et al. Illumina TruPath genome towards comprehensive high-resolution analysis of genomic diversity and structural complexity. Npj Genomic Med. 2026 Sep 1. doi:10.1038/s41525-026-00610-y

23. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023;21:2352–64. doi:10.1016/j.csbj.2023.03.038 PubMed PMID: 37025654; PubMed Central PMCID: PMC10070092.

24. MacKenzie M, Argyropoulos C. An Introduction to Nanopore Sequencing: Past, Present, and Future Considerations. Micromachines. 2023 Feb 16;14(2):459. doi:10.3390/mi14020459 PubMed PMID: 36838159; PubMed Central PMCID: PMC9966803.

25. Goenka SD, Gorzynski JE, Shafin K, Fisk DG, Pesout T, Jensen TD, et al. Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. Nat Biotechnol. 2022 Jul;40(7):1035–41. doi:10.1038/s41587-022-01221-5 PubMed PMID: 35347328; PubMed Central PMCID: PMC9287171.

26. White LK, Hesselberth JR. Modification mapping by nanopore sequencing. Front Genet. 2022;13:1037134. doi:10.3389/fgene.2022.1037134 PubMed PMID: 36386798; PubMed Central PMCID: PMC9650216.

27. Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, et al. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. Genome Res. 2023 Dec 27;33(12):2029–40. doi:10.1101/gr.278070.123 PubMed PMID: 38190646; PubMed Central PMCID: PMC10760522.

28. Eveleigh RJM, Reiling SJ, Galvez JH, Bourgey M, Ragoussis J, Bourque G. Benchmarking of sequencing technologies defines optimal strategies for genetic variants detection in a human genome. Genome Biol. 2026 Apr 8;27(1):129. doi:10.1186/s13059-026-04048-4 PubMed PMID: 41952156; PubMed Central PMCID: PMC13072666.

29. Zhou L, Ng HK, Drautz-Moses DI, Schuster SC, Beck S, Kim C, et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. Sci Rep. 2019 Jul 17;9(1):10383. doi:10.1038/s41598-019-46875-5 PubMed PMID: 31316107; PubMed Central PMCID: PMC6637168.

30. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, et al. Benchmarking challenging small variants with linked and long reads. Cell Genomics. 2022 May;2(5):100128. doi:10.1016/j.xgen.2022.100128 PubMed PMID: 36452119; PubMed Central PMCID: PMC9706577.

31. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics. 2022 May 11;2(5):100129, S2666-979X(22)00058-1. doi:10.1016/j.xgen.2022.100129 PubMed PMID: 35720974; PubMed Central PMCID: PMC9205427.

32. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013 Oct 10;155(2):462–77. doi:10.1016/j.cell.2013.09.034 PubMed PMID: 24120142; PubMed Central PMCID: PMC3910500.

33. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061–8. doi:10.1038/nature07385 PubMed PMID: 18772890; PubMed Central PMCID: PMC2671642.

34. Koelsche C, Schrimpf D, Stichel D, Sill M, Sahm F, Reuss DE, et al. Sarcoma classification by DNA methylation profiling. Nat Commun. 2021 Jan 21;12(1):498. doi:10.1038/s41467-020-20603-4 PubMed PMID: 33479225; PubMed Central PMCID: PMC7819999.

35. de Abreu AR, Ibrahim J, Lemonidis V, Mateiu L, Van Camp G, Op de Beeck K. Comparison of current methods for genome-wide DNA methylation profiling. Epigenetics Chromatin. 2025 Aug 25;18(1):57. doi:10.1186/s13072-025-00616-3 PubMed PMID: 40855329; PubMed Central PMCID: PMC12376410.

36. Guanzon D, Ross JP, Ma C, Berry O, Liew YJ. Comparing methylation levels assayed in GC-rich regions with current and emerging methods. BMC Genomics. 2024 Jul 30;25(1):741. doi:10.1186/s12864-024-10605-7 PubMed PMID: 39080541; PubMed Central PMCID: PMC11289974.

37. Guo X, Chen Q, Zhang Y, Zhang Y, Liu Y, Duan S, et al. Methylation reference datasets from quartet DNA materials for benchmarking epigenome sequencing. Nat Commun. 2025 Oct 16;16(1):9202. doi:10.1038/s41467-025-64250-z PubMed PMID: 41102181; PubMed Central PMCID: PMC12533135.

38. Simons RB, Karkala F, Kukk MM, Adams HHH, Kayser M, Vidaki A. Comparative performance evaluation of bisulfite- and enzyme-based DNA conversion methods. Clin Epigenetics. 2025 Apr 3;17(1):56. doi:10.1186/s13148-025-01855-7 PubMed PMID: 40181442; PubMed Central PMCID: PMC11969950.

39. Kulkarni O, Mathew RJ, Jana R, Zaveri L, Ara S, Nagabandi T, et al. Comprehensive benchmarking of tools for nanopore-based detection of DNA methylation. Nat Commun. 2026 Jul 15;17(1):8685. doi:10.1038/s41467-026-75183-6 PubMed PMID: 42457719; PubMed Central PMCID: PMC13490465.

40. Dwarshuis N, Kalra D, McDaniel J, Sanio P, Alvarez Jerez P, Jadhav B, et al. The GIAB genomic stratifications resource for human reference genomes. Nat Commun. 2024 Oct 19;15(1):9029. doi:10.1038/s41467-024-53260-y PubMed PMID: 39424793; PubMed Central PMCID: PMC11489684.

41. Rocco JW. Mutant allele tumor heterogeneity (MATH) and head and neck squamous cell carcinoma. Head Neck Pathol. 2015 Mar;9(1):1–5. doi:10.1007/s12105-015-0617-1 PubMed PMID: 25804377; PubMed Central PMCID: PMC4382477.

42. Wu P, Yang W, Ma J, Zhang J, Liao M, Xu L, et al. Mutant-allele tumor heterogeneity in malignant glioma effectively predicts neoplastic recurrence. Oncol Lett. 2019 Dec;18(6):6108–16. doi:10.3892/ol.2019.10978 PubMed PMID: 31788085; PubMed Central PMCID: PMC6865645.

43. Kinnersley B, Jung J, Cornish AJ, Chubb D, Laxton R, Frangou A, et al. Genomic landscape of diffuse glioma revealed by whole genome sequencing. Nat Commun. 2025 May 7;16(1):4233. doi:10.1038/s41467-025-59156-9 PubMed PMID: 40335506; PubMed Central PMCID: PMC12059081.

44. Stichel D, Ebrahimi A, Reuss D, Schrimpf D, Ono T, Shirahata M, et al. Distribution of EGFR amplification, combined chromosome 7 gain and chromosome 10 loss, and TERT promoter mutation in brain tumors and their potential for the reclassification of IDHwt astrocytoma to glioblastoma. Acta Neuropathol (Berl). 2018 Nov;136(5):793–803. doi:10.1007/s00401-018-1905-0 PubMed PMID: 30187121.

45. Daenekas B, Pérez E, Boniolo F, Stefan S, Benfatto S, Sill M, et al. Conumee 2.0: enhanced copy-number variation analysis from DNA methylation arrays for humans and mice. Bioinformatics. 2024 Feb 1;40(2):btae029. doi:10.1093/bioinformatics/btae029 PubMed PMID: 38244574; PubMed Central PMCID: PMC10868300.

46. Vasan N, Razavi P, Johnson JL, Shao H, Shah H, Antoine A, et al. Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kα inhibitors. Science. 2019 Nov 8;366(6466):714–23. doi:10.1126/science.aaw9032 PubMed PMID: 31699932; PubMed Central PMCID: PMC7173400.

47. Akbari V, Jones SJM. Phasing DNA Methylation. Methods Mol Biol. 2023;2590:219–35. doi:10.1007/978-1-0716-2819-5_14 PubMed PMID: 36335502.

48. Martin M, Ebert P, Marschall T. Read-Based Phasing and Analysis of Phased Variants with WhatsHap. Methods Mol Biol. 2023;2590:127–38. doi:10.1007/978-1-0716-2819-5_8 PubMed PMID: 36335496.

49. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005 Mar 10;352(10):997–1003. doi:10.1056/NEJMoa043331 PubMed PMID: 15758010.

50. Li Q, Keskus AG, Wagner J, Izydorczyk MB, Timp W, Sedlazeck FJ, et al. Unraveling the hidden complexity of cancer through long-read sequencing. Genome Res. 2025 Apr 14;35(4):599–620. doi:10.1101/gr.280041.124 PubMed PMID: 40113261; PubMed Central PMCID: PMC12047254.

51. Warburton PE, Sebra RP. Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. Annu Rev Genomics Hum Genet. 2023 Aug 25;24:109–32. doi:10.1146/annurev-genom-101722-103045 PubMed PMID: 37075062.

52. Sarashetti P, Lipovac J, Tomas F, Šikić M, Liu J. Evaluating data requirements for high-quality haplotype-resolved genomes for creating robust pangenome references. Genome Biol. 2024 Dec 18;25(1):312. doi:10.1186/s13059-024-03452-y PubMed PMID: 39696427; PubMed Central PMCID: PMC11658127.

53. Chen L, Zheng Z, Su J, Yu X, Wong AOK, Zhang J, et al. ClairS-TO: a deep-learning method for long-read tumor-only somatic small variant calling. Nat Commun. 2025 Oct 31;16(1):9630. doi:10.1038/s41467-025-64547-z PubMed PMID: 41173866; PubMed Central PMCID: PMC12579226.

54. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. 2018 Nov 16;46(20):e123. doi:10.1093/nar/gky691 PubMed PMID: 30085201; PubMed Central PMCID: PMC6237738.

55. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987–93. doi:10.1093/bioinformatics/btr509 PubMed PMID: 21903627; PubMed Central PMCID: PMC3198575.

56. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6;17(1):122. doi:10.1186/s13059-016-0974-4 PubMed PMID: 27268795; PubMed Central PMCID: PMC4893825.

57. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018 Nov;28(11):1747–56. doi:10.1101/gr.239244.118 PubMed PMID: 30341162; PubMed Central PMCID: PMC6211645.

58. Elrick H, Sauer CM, Espejo Valle-Inclan J, Trevers K, Tanguy M, Zumalave S, et al. SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing. Nat Methods. 2025 Jul;22(7):1436–46. doi:10.1038/s41592-025-02708-0 PubMed PMID: 40437218; PubMed Central PMCID: PMC12240814.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pptx
- SupplementaryTables.xlsx