Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P Kuschel, Jürgen Hench, Stephan Frank, Ivana Bratic Hench, Elodie Girard, Maud Blanluet, Julien Masliah-Planchon, Martin Misch, Julia Onken, Marcus Czabanka, et al.

Received: 3 June 2021 | Revised: 28 August 2022 | Accepted: 2 October 2022

DOI: 10.1111/nan.12856

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology
JOURNAL OF THE BRITISH NEUROPATHOLOGICAL SOCIETY

WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

[1]Department of Neurology with Experimental Neurology, Charité – Universitätsmedizin Berlin, Berlin, Germany

[2]Department of Pathology, Universitätsspital Basel, Basel, Switzerland

[3]Institut Curie, Paris, France

[4]Department of Neurosurgery, Charité – Universitätsmedizin Berlin, Berlin, Germany

[5]Center for Digital Health, Berlin Institute of Health (BIH) and Charité - Universitätsmedizin Berlin, Berlin, Germany

[6]Department of Neuropathology, Charité – Universitätsmedizin Berlin, Berlin, Germany

[7]Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neurologie 2-Mazarin, F-75013, Paris, France

[8]Center for Stroke Research, Berlin, Germany

[9]German Cancer Consortium (DKTK), Partner Site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany

Correspondence
Philipp Euskirchen, Department of Neurology with Experimental Neurology, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.
Email: philipp.euskirchen@charite.de

Funding information
The study was funded by the Brain Tumour Charity (UK) (GN-000694).

Abstract

Background: DNA methylation-based classification of cancer provides a comprehensive molecular approach to diagnose tumours. In fact, DNA methylation profiling of human brain tumours already profoundly impacts clinical neuro-oncology. However, current implementation using hybridisation microarrays is time consuming and costly. We recently reported on shallow nanopore whole-genome sequencing for rapid and cost-effective generation of genome-wide 5-methylcytosine profiles as input to supervised classification. Here, we demonstrate that this approach allows us to discriminate a wide spectrum of primary brain tumours.

Results: Using public reference data of 82 distinct tumour entities, we performed nanopore genome sequencing on 382 tissue samples covering 46 brain tumour (sub)types. Using bootstrap sampling in a cohort of 55 cases, we found that a minimum set of 1000 random CpG features is sufficient for high-confidence classification by ad hoc random forests. We implemented score recalibration as a confidence measure for interpretation in a clinical context and empirically determined a platform-specific threshold in a randomly sampled discovery cohort ($N = 185$). Applying this cut-off to an independent validation series ($n = 184$) yielded 148 classifiable cases (sensitivity 80.4%) and demonstrated 100% specificity. Cross-lab validation demonstrated robustness with concordant results across four laboratories in 10/11 (90.9%) cases. In a prospective benchmarking ($N = 15$), the median time to results was 21.1 h.

Conclusions: In conclusion, nanopore sequencing allows robust and rapid methylation-based classification across the full spectrum of brain tumours. Platform-specific confidence scores facilitate clinical implementation for which prospective evaluation is warranted and ongoing.

KEYWORDS
brain tumour, epigenomics, machine learning, molecular pathology, nanopore sequencing, whole-genome sequencing

BACKGROUND

DNA methylation is a stable epigenomic mark of cell identity and has become a powerful tool in tissue-based cancer diagnosis. The application of supervised machine learning to genome-wide DNA methylation profiles enables the classification of unknown samples with respect to a reference cohort, providing a method that is both independent and complementary to histomorphology. It has been used in a variety of clinical scenarios, such as classification of brain tumours [1], soft tissue sarcoma [2] or cancer of unknown primary (CUP) [3]. In addition to offering an unbiased method of mandatory molecular testing, required by the current WHO classification of CNS tumours (e.g., for medulloblastoma subtypes), systematic application of methylation-based classification has revealed weaknesses of histomorphology, which is associated with misdiagnoses in approximately 10% of cases [1]. However, for current microarray-based implementation, turnaround times are in the range of several days to weeks [4] due to complex wet-lab procedures and sample multiplexing. Together with high capital costs, reasonable per-assay costs are only achieved in high-throughput settings like tertiary care centres.

To address these shortcomings, we have recently used nanopore whole-genome sequencing (WGS) for the simultaneous generation of copy number and DNA methylation profiles [5]. Nanopore sequencing determines DNA sequence as well as base modifications, such as 5-methylation of cytosine (5mC), by detecting changes in ionic currents when molecules pass a biological nanopore inserted in a dielectric membrane [6, 7]. Methylation detection in native DNA, in combination with time-efficient hands-on procedures in the range of less than an hour in total, significantly decreases turnaround times.

Here, we extend this pilot approach to methylation-based classification to 82 brain tumour entities, define criteria for robust diagnostic implementation and provide benchmarking data from cross-laboratory as well as cross-method validation. Moreover, we share first insights into the prospective analysis of turnaround times in comparison to routine diagnostic procedures as part of an ongoing multicentric clinical trial.

Key Points

- Nanopore methylome sequencing allows for high-confidence methylation-based classification of brain tumours.
- A platform-specific threshold for classification scores results in high specificity while maintaining sensitivity.
- Easy setup, low per-assay cost and rapid turnaround times enable widespread implementation.

METHODS

Experimental design

A total of 382 brain tumour biopsies were included in this study, comprising 15 prospective cases, 41 retrospective cases from previously published datasets [5, 8] and 326 cases collected during routine clinical application at the Division of Neuropathology, Institute of Pathology, Basel, Switzerland (as an in-house validated diagnostic test) (Figure S1). The local ethics committee (Charité – Universitätsmedizin Berlin, Berlin, Germany; EA2/041/18) approved generation of prospective data in the context of this study. In order to refine machine learning and to define quality criteria for reliable classification, the cohort was randomly split (50/50) into a discovery and validation cohort ($N = 191$ each). All cases underwent routine diagnostic procedures, including microarray-based analysis [1] in 321/382 (84.03%) cases, and were classified in accordance with the 2016 World Health Organisation Classification of Tumours of the Central Nervous System [9] (Table 1).

Patient material and tissue processing

Fresh tumour tissue of prospective cases was transferred to the local neuropathology laboratory during routine diagnostic procedures. Tissue was then snap-frozen, and H&E cryosections were inspected to assess tumour purity. DNA was extracted using spin columns (DNeasy Blood & Tissue Kit, Qiagen, NL) according to the manufacturer's protocol with ~25 mg of tumour tissue. For the cross-laboratory cohort, tumour DNA obtained from the Institute Curie was extracted using phenol/chloroform. Eluted genomic DNA was quantified on a Qubit 4.0 fluorometer using the dsDNA BR Assay (Thermo Fisher, USA) and quality controlled using the 260/280 ratio (NanoDrop, Thermo Fisher, USA).

Nanopore WGS

Library preparation with barcode labelling was performed with ~400 ng input of genomic DNA using the Rapid Barcoding Kit (RBK004, Oxford Nanopore Technologies, UK) according to the manufacturer's instructions. During the library preparation, input DNA is fragmented while simultaneously attaching barcodes using a time-efficient transposase-based approach. The final library was loaded onto an R9.4.1 flow cell (FLO-MIN106D, Oxford Nanopore Technologies, UK; alternatively, FLG-0001 cells sharing the architecture with FLO-MIN106D were used), and WGS was performed for 6–24 h on a MinION Mk 1B device (Oxford Nanopore Technologies, UK). FAST5 files

TABLE 1 Summary of tumour entities in this study

	Discovery (*N* = 191)	Validation (*N* = 191)	Overall (*N* = 382)
WHO 2016 integrated diagnosis			
Adamantinomatous craniopharyngioma	1 (0.5%)	3 (1.6%)	4 (1.0%)
Anaplastic astrocytoma, IDH-mutant	7 (3.7%)	7 (3.7%)	14 (3.7%)
Anaplastic oligodendroglioma, IDH-mutant and 1p/19q-codeleted	5 (2.6%)	3 (1.6%)	8 (2.1%)
Atypical teratoid/rhabdoid tumour	2 (1.0%)	1 (0.5%)	3 (0.8%)
Central neurocytoma	1 (0.5%)	0 (0%)	1 (0.3%)
Chordoma	1 (0.5%)	1 (0.5%)	2 (0.5%)
CNS neuroblastoma with FOXR2 activation	1 (0.5%)	0 (0%)	1 (0.3%)
Diffuse astrocytoma, IDH-mutant	2 (1.0%)	1 (0.5%)	3 (0.8%)
Diffuse large B-cell lymphoma (DLBCL)	5 (2.6%)	4 (2.1%)	9 (2.4%)
Diffuse leptomeningeal glioneuronal tumour	1 (0.5%)	0 (0%)	1 (0.3%)
Diffuse midline glioma, H3 K27M-mutant	1 (0.5%)	0 (0%)	1 (0.3%)
Embryonal tumour with multilayered rosettes, C19MC-altered	1 (0.5%)	2 (1.0%)	3 (0.8%)
Ependymoma	4 (2.1%)	2 (1.0%)	6 (1.6%)
Glioblastoma, IDH wild type, H3.3 G34 mutant[a]	1 (0.5%)	0 (0%)	1 (0.3%)
Glioblastoma, IDH wild type, subclass midline[a]	1 (0.5%)	0 (0%)	1 (0.3%)
Glioblastoma, IDH-mutant	3 (1.6%)	2 (1.0%)	5 (1.3%)
Glioblastoma, IDH-wild type	48 (25.1%)	52 (27.2%)	100 (26.2%)
Haemangioblastoma	1 (0.5%)	1 (0.5%)	2 (0.5%)
Medulloblastoma, genetically defined, group 3	1 (0.5%)	3 (1.6%)	4 (1.0%)
Medulloblastoma, genetically defined, non-WNT/non-SHH	1 (0.5%)	3 (1.6%)	4 (1.0%)
Medulloblastoma, genetically defined, SHH-activated and TP53-wildtype	1 (0.5%)	0 (0%)	1 (0.3%)
Medulloblastoma, genetically defined, WNT-activated	1 (0.5%)	1 (0.5%)	2 (0.5%)
Meningioma	60 (31.4%)	62 (32.5%)	122 (31.9%)
Pilocytic astrocytoma	6 (3.1%)	4 (2.1%)	10 (2.6%)
Pituitary adenoma ACTH producing	2 (1.0%)	4 (2.1%)	6 (1.6%)
Pituitary adenoma densely granulated GH/STH producing	3 (1.6%)	1 (0.5%)	4 (1.0%)
Pituitary adenoma gonadotropin producing	16 (8.4%)	10 (5.2%)	26 (6.8%)
Pituitary adenoma sparsely granulated GH/STH producing	1 (0.5%)	1 (0.5%)	2 (0.5%)
Pituitary adenoma TSH producing	1 (0.5%)	0 (0%)	1 (0.3%)
Rosette-forming glioneuronal tumour	1 (0.5%)	0 (0%)	1 (0.3%)
Schwannoma	10 (5.2%)	12 (6.3%)	22 (5.8%)
Solitary fibrous tumour/haemangiopericytoma	1 (0.5%)	1 (0.5%)	2 (0.5%)
Anaplastic pilocytic astrocytoma	0 (0%)	1 (0.5%)	1 (0.3%)
CNS embryonal tumour, NOS	0 (0%)	1 (0.5%)	1 (0.3%)
CNS Ewing sarcoma family tumour with CIC alteration	0 (0%)	1 (0.5%)	1 (0.3%)
Medulloblastoma, NOS	0 (0%)	1 (0.5%)	1 (0.3%)
Medulloblastoma, SHH-activated	0 (0%)	1 (0.5%)	1 (0.3%)
Pleomorphic xanthoastrocytoma	0 (0%)	2 (1.0%)	2 (0.5%)

(Continues)

TABLE 1 (Continued)

	Discovery ($N = 191$)	Validation ($N = 191$)	Overall ($N = 382$)
Subependymal giant cell astrocytoma	0 (0%)	2 (1.0%)	2 (0.5%)
Subependymoma	0 (0%)	1 (0.5%)	1 (0.3%)

Reference diagnosis is reported in accordance with the 2016 WHO classification of central nervous system tumours. A detailed summary of the clinical characteristics of all cases can be found in Table S1.

[a]Of note, these entities are not yet recognised as distinct entities in the 2016 WHO classification.

containing the raw data were obtained in real time using the manufacturer's software MinKNOW (v.1.3.1-v.3.6.0) and transferred to a high-performance computing (HPC) cluster for further analysis. Each flow cell was washed after sequencing (WSH002/WSH003, Oxford Nanopore Technologies, UK) and reused for up to four samples. When multiplexing retrospective samples, up to five libraries were sequenced simultaneously, and sequencing for up to 24 h was performed.

Sequencing data processing

Base calling of raw data was performed using the manufacturer's proprietary software (guppy v3.1.5, CPU-based, fast mode, or v3.4.3, GPU-based, Oxford Nanopore Technologies, UK). Demultiplexing to identify carry-over of barcodes from previous samples run on the same flow cell was performed. No significant (> 1%) cross-contamination was detected in any sample, and all reads were used for subsequent analysis. Reads were then aligned to the hg19 human reference genome using minimap2 v.2.15 [10]. Copy number profiles were generated using R/Bioconductor and the QDNAseq package v.1.20 [11] using public data from a single flow cell sequencing run (FAF04090) generated with NA12878 reference DNA [12] for pseudo-germline subtraction.

The methylation status of CpG sites (5mC) was called using nanopolish v0.13.2 [7]. All workflows were implemented using snakemake v5.4.0–v.6.1.1 [13] for parallelisation and deployment to an HPC cluster. When not using HPC, the pipeline was deployed to a single x86_64 workstation with 120 cores and 2 TB RAM, augmented with an RTX2070 consumer-grade GPU (NVIDIA, Santa Clara, CA, USA) for base calling, running a Linux operating system (Ubuntu 18.04).

Random forest classification

5mC signals at sites overlapping with sites probed by the Illumina BeadChip 450 K array were then used to train a random forest (RF) classifier using the Heidelberg reference cohort of brain tumour methylation profiles [1]. Thirteen out of 382 samples (3.4%) with less than 1000 overlapping CpG sites between nanopore and reference data were excluded to ensure a minimum technical quality of sequencing data generated. Beta values from the training set and sample set were binarised using 0.6 as a threshold value. When more than 50,000 features, the 50,000 most variable features were selected by standard deviation. RF classification was implemented in Python using the RandomForestClassifier from the scikit-learn package v.1.0.2 [14]. The following parameters were modified from default: n_estimators = 5000. Stratified sampling (to match the smallest class size, i.e. eight) was used to account for class imbalance. Although RF are effective for many classification tasks, they typically report poor estimates of class probabilities. Given the importance of interpretable probability scores in a clinical context, the estimated class probabilities from such a classifier can be calibrated, which rescales predicted probabilities to be more accurately interpreted as confidence levels. We adopted a calibration strategy using Platt scaling by fitting a logistic regression model as previously described [15] as follows. In order to recalibrate the output of the RF classifier, we use fivefold cross-validation to collect the raw prediction probabilities from the independent cross-validation fold. The collection of raw prediction probabilities was then used to train the calibration model. Platt scaling was originally suggested for binary classification tasks, so we reduced our multi-class classification task to a series of binary calibration tasks using the 1-vs-rest method [16]. Platt scaling was implemented by fitting the sigmoid regression using the CalibratedClassifierCV function from the scikit-learn package in Python [14].

Prospective cases and cross-laboratory testing

As a pilot phase for a multicentric clinical trial (Universal Trial Number: U1111-1239-3456), a small series of prospective cases ($N = 15$) were evaluated regarding turnaround time and concordance to results from routine diagnostic workup. Prospective samples were obtained from the local neurosurgery department after written informed consent was given by the patient. Pseudonymised study data (start of surgery, time point of tissue receipt, DNA extraction, library preparation, sequencing metrics, classification results) was collected and managed using REDCap software [17], which was provided by the Berlin Institute of Health's Clinical Research Unit in a certified computing environment.

Statistics

Data analysis was performed using R v.4.0.2. Receiver operator characteristics (ROC) were analysed with the ROCit v.2.1.1 package. Figures were mainly visualised using ggplot2 v.3.3.2.

Code and data availability

The current nanoDx classification and analysis pipeline is publicly available at https://gitlab.com/pesk/nanoDx (version v.0.4.0rc1 was used for pre-processing of all sequencing data). The source code for the outlined RF implementation and to reproduce all analyses and figures in this manuscript is available at https://gitlab.com/pesk/nanoBenchmark. Raw sequencing data from 56 samples have been deposited at the European Genome-Phenome Archive (EGAS00001006540 and EGAS00001002213). Methylation microarray raw data and methylation calls are deposited at Gene Expression Omnibus (GSE209865).

RESULTS

Robustness of pan-brain cancer classification using nanopore sequencing

In low-pass nanopore WGS, genome coverage is sparse, and only a random subset of the ~30 million CpG sites in the genome are probed. We have therefore proposed ad hoc training of RF using the overlap of the random CpG feature set of each sequencing run and the fixed feature space of the microarray-based training set [5]. Here, we used the Heidelberg brain tumour classifier as a reference, which distinguishes 82 tumour entities and nine non-tumour control classes. It implements a two-tier approach for methylation classes that are prone to misclassification without current clinical consequences (e.g. subtypes of IDH-wild-type glioblastoma). In these situations, superclasses, termed methylation class families (MCFs), were defined followed by subtype classification where applicable. Similarly, we performed classification against the full training set of 91 methylation classes. For MCF level classification (composed of eight MCFs and 67 methylation classes, totalling 75 classes), the resulting recalibrated votes by methylation classes were grouped by their respective MCFs (if available), and the addition of recalibrated votes was performed. Classification results were compared to the institutional WHO 2016 integrated diagnosis. The majority vote across the discovery cohort of 185 brain tumour samples was correct in 172/185 (93.0%) and 177/185 (95.7%) cases for the MCF and full training set, respectively.

As a measure of certainty of the classification, we recalibrated scores by Platt scaling of RF raw votes. This transformation allows the application of a single cut-off value to identify valid predictions. We used ROC analysis to identify the optimal cut-off value in the discovery cohort. For MCF level classification (75 classes), the calibrated score allowed reliable prediction of correct classification (AUC = 0.95; Figure 1A) by applying a cut-off value of >0.15 (which is slightly more conservative than the optimal cut-off of >0.13 suggested by Youden index interpretation). This resulted in 86.5% sensitivity and 100% specificity in the discovery cohort (Figure 1C). Youden index analysis suggested a similar cut-off (>0.13) for the full methylation class-level (91 classes) training set (AUC = 0.98; Figure 1B). Therefore, we decided to implement >0.15 as an empirical nanopore-specific threshold for both classification methods. This resulted in 75.7% sensitivity and 100% specificity to correctly predict methylation class in the discovery cohort (Figure S2A).

We then applied this approach to the independent validation cohort of 184 primary brain tumour cases. On the MCF level, 167/184 cases (90.7%) were classified correctly overall, that is, a methylation class concordant with the sample's WHO integrated diagnosis was called. Applying the nanopore-specific threshold >0.15 resulted in the correct classification of all 148 cases with scores above the cut-off, corresponding to 80.4% sensitivity and 100% specificity (Figure 1C,D). On the methylation class level, overall correct classification was found in 170/184 cases (92.4%). Requiring a cut-off >0.15 yielded 69.7% sensitivity while retaining 100% specificity (Figure S2A,B).

Matched array-based methylation data and classification results were available for 312/369 (84.6%) cases. The microarray data IDAT files were uploaded to the Heidelberg classifier, and the methylation class output was compared to our nanopore-based approach. Identical methylation classes were assigned in 254/312 (81.4%) cases overall and in 214/232 (92.2%) cases where a nanopore sequencing-based classification yielded a score >0.15 (Table S1). All discordant cases 12/232 (5.2%) were IDH-wild-type glioblastomas, and classification was concordant at the MCF level, but different glioblastoma subtypes were called.

The impact of the number of CpG features

Next, we investigated whether there was a correlation between the number of CpG features and recalibrated classification score, as we hypothesised that a minimum number of features would be required for robust classification. Surprisingly, no significant correlation between the number of CpG features and the classification score was observed (Pearson's $r = 0.06$, $p = 0.4$) within the discovery cohort. Next, we performed iterative random subsampling of CpG features for a cohort of 55 samples with five different seeds. However, when less than 1000 CpG sites were used, low scores were frequent, and even misclassification with high confidence scores was observed (Figure S3). Additionally, when investigating all subsampling iterations with a minimum of 1000 CpGs and a score above 0.15, the correct call rate was 1289/1290 (99.9%). We, therefore, considered 1000 overlapping CpG sites between the fixed feature space of the microarray-based training set and whole-genome nanopore sequencing, the absolute minimum number for reasonable analysis.

Cross-laboratory validation

Next, to assess reproducibility across laboratories, a series of 11 tumours were profiled independently by nanopore WGS at four sites using the same DNA sample. Global correlation of all methylation calls using Pearson's *r* between any two samples reproducibly showed the best correlation for matched pairs (Figure 2A). Classification of matched samples was identical in 10/11 (90.1%) cases (Table S2) with very similar raw vote distributions (Figure 2B). In addition, DNA

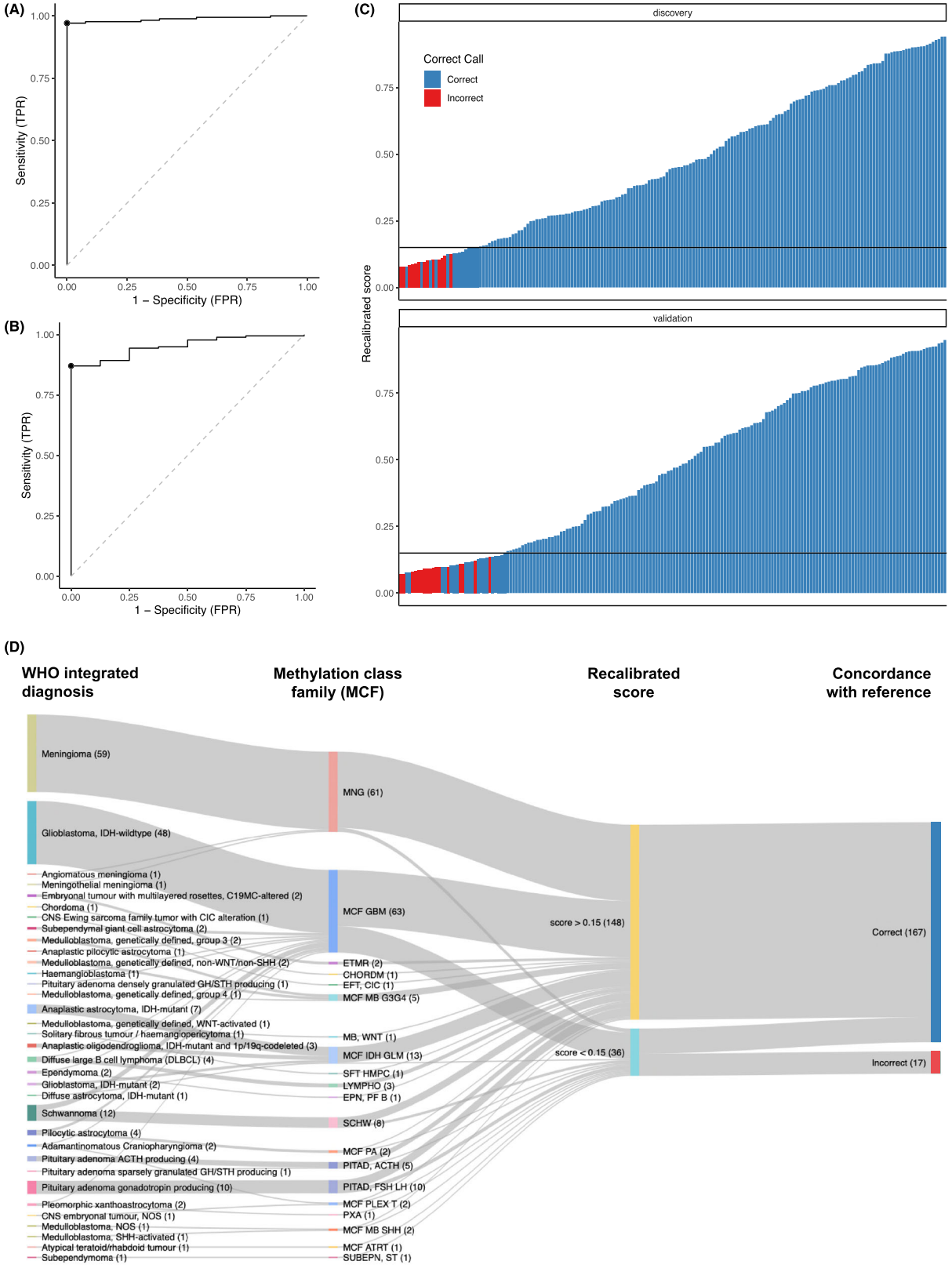

FIGURE 1 Legend on next page.

FIGURE 1 Performance of ad hoc random forest pan-brain tumour classification. (A,B) Determination of cut-offs for recalibrated scores to detect correct classification using a methylation class family (A) and methylation class (B) training set. (C) Waterfall plots indicate the relation of the methylation class family level classification result and the recalibrated classification score in the discovery ($N = 185$) and validation cohort ($N = 184$). Colour indicates concordance (*blue*) or discordance (*red*) of the called methylation class family or methylation class with the institutional WHO 2016 integrated diagnosis. The cut-off identified by ROC analysis (>0.15) is indicated by the horizontal solid line. (D) Classification results with respect to WHO diagnosis and corresponding methylation classification result in the validation cohort of $N = 184$ independent samples.

methylation microarray (Illumina Infinium BeadArray 850 K chip) data were available for all samples. Concordant results across all nanopore and microarray-based tests were obtained in 10/11 (90.9%) cases. In the one mismatch case, the nanopore-based classification matched the histology-only based reference diagnosis (medulloblastoma, NOS) in 2/4 (50.0%) cases on MCF level and 3/4 cases (75.0%) on MC level. The corresponding microarray resulted in classification as non-neoplastic (cerebellar) control.

Benchmarking of nanopore sequencing over time

Nanopore sequencing data are generated sequentially over time and permits real-time analysis. Therefore, next, we reanalysed a cohort of 56 tumours by extracting subsets of sequencing data that were generated within a given time interval. As the raw yield is roughly proportional to elapsed sequencing time, the number of sampled CpG sites overlapping with the training set steadily increased over time. The minimum number of 1000 CpG sites was sequenced within the first 30 min of the nanopore run in 38/56 samples (67.9%; Figure 3B). After 90 min, the correct call rate from cases with a minimum of 1000 and a score above 0.15 with later correct classification result with their full set of CpG features was 100% (Figure 3A). A correct, high-confidence classification (i.e. correct call with score > 0.15) was made within 30 min of sequencing in 38/56 samples (67.9%). Importantly, high-confidence classifications based upon more than 1000 CpG features, no matter at what time point they were made, were correct in 54/56 (96.4%) samples (Figure S4).

Benchmarking of turnaround times

Finally, as part of the pilot phase for a multicentric clinical trial, samples from 15 patients were analysed prospectively. Results of nanopore methylation-based classification were available in a mean of 39.4 h (median = 21.1 h) after receiving tissue from histological tumour purity assessment (Figure 3C). Furthermore, hands-on time spent on DNA extraction, quality control and nanopore library preparation was 146.5 min on average for singleplex sequencing (Figure 3D).

DISCUSSION

In the present study, we demonstrate that methylation profiles generated by low-pass nanopore WGS allow robust and unbiased classification of primary brain tumours over the entire spectrum covered by the Heidelberg brain tumour classifier. We define cut-off values for reliable clinical interpretation and show that ad hoc training and classification using RF yields classification with very high specificity and acceptable sensitivity in real-world data. Moreover, results are concordant and reproducible across laboratories. The sensitivity of the method on MCF and MC levels was 80.4% and 69.7%, respectively, for the >0.15 cut-off. This is comparable to previously reported sensitivity for microarray-based classification, which ranged from 56% in a real-world cohort enriched for challenging cases [4] to 88% in a well-defined validation cohort [1].

Opportunities for nanopore methylation-based tumour classification

With a median time to diagnosis of 21 h within the prospective cohort, nanopore-based methylation classification offers the chance to dramatically shorten turnaround times allowing completion of methylation-based molecular profiling without delaying first-line therapy. Nanopore methylation-based classification was usually available prior to immunohistochemistry in routine pathological workup due to delay by fixation and paraffin embedding. This allows neuropathologists to decide which staining and complementary methods could further contribute to the diagnostic procedure, avoiding time-consuming sequential testing and accelerating the overall time to integrated diagnosis. Moreover, our data suggest a further potential to reduce the turnaround time of nanopore-based testing, and even intraoperative classification may be possible. In addition, genome-wide copy number profiles generated from nanopore WGS data have proven highly useful in diagnostic decision making, replacing, for example, time-consuming and costly fluorescence in situ hybridisation assays.

Machine learning aspects

Our current implementation of ad hoc random forests uses binarisation of methylated allele frequencies to normalise for platform differences between microarrays (reference data) and nanopore sequencing. Consequently, the binominal distribution of CpG methylation due to low coverage of nanopore data results in significantly lower recalibrated scores compared to the corresponding microarray-based classification score. By implementing a platform-specific threshold, we account for this and allow reliable interpretation in clinical applications. However, recalibrated scores cannot be interpreted as

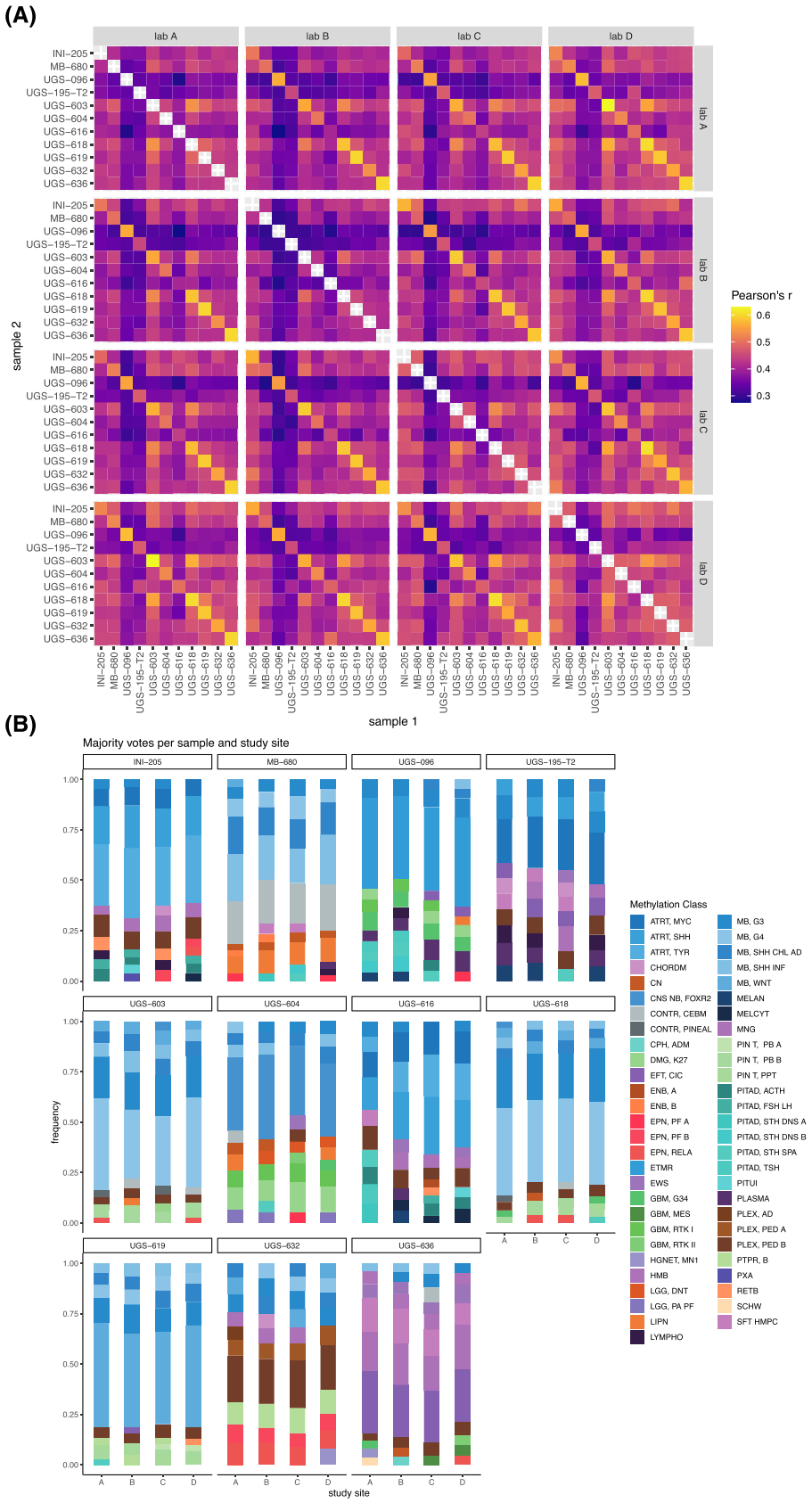

FIGURE 2 Cross-laboratory validation. (A) Pearson's correlation of methylation status of shared CpG features between paired samples of four laboratory sites. (B) Comparison of random forest vote distribution for each sample quadruplet. Bar plots show the Top 10 raw majority votes per methylation class.

probabilities. Resampling simulated binary distributions from the microarray reference data as input for training the model may further improve interpretability in the future. In addition, CpG sites called from individual long nanopore reads are not independent features but are currently treated this way. Exploiting this epi-haplotype information could possibly further improve the classification.

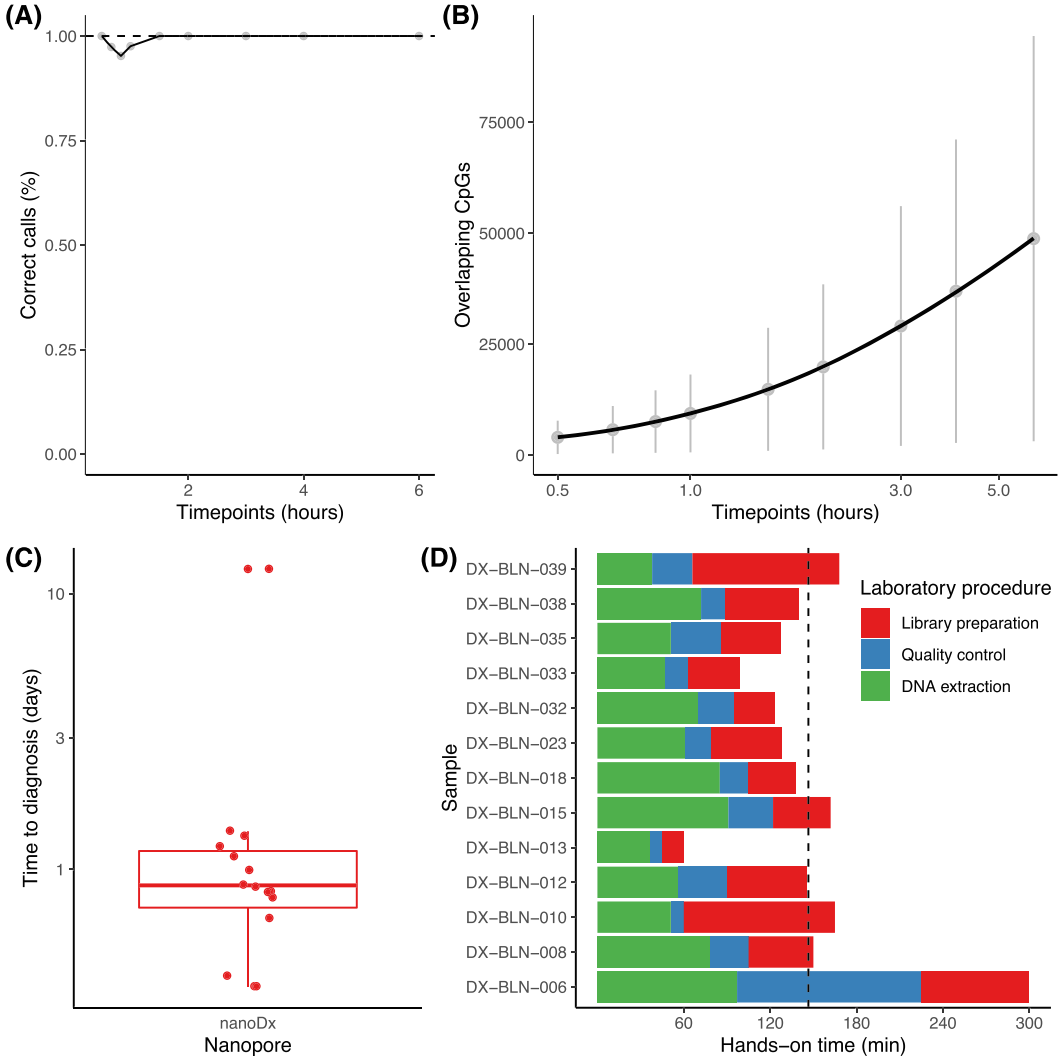

FIGURE 3 Benchmarking of nanopore sequencing over time and laboratory turnaround times. (A) Per cent of correctly classified samples in the discovery cohort with a score >0.15 with a correct majority vote at a given time point (in hours). (B) Number of CpG features with respect to sequencing time passed. Light grey lines show the standard deviation for each time point. (C) Boxplots indicate the median and quartile ranges of time elapsed from receipt of tumour material to timestamp of nanopore sequencing-based classification reports, respectively. (D) Analysis of hands-on time with respect to time spent on DNA extraction (*green*), DNA quality control and quantification (*blue*) and library preparation (*red*). The average time elapsed across all samples is indicated as a dashed line.

Limitations

Although this study aims to offer benchmarking of nanopore sequencing for DNA methylation-based classification of brain tumours and demonstrates its robustness, further research is needed in order to evaluate non-inferiority against current gold standard techniques (e.g. methylation bead array techniques) in larger cohorts. This should especially be evaluated regarding turnaround times with different setup and computing options. This is the scope of an ongoing multicentric clinical trial (German Clinical Trials Register, Universal Trial ID: U1111-1239-3456). In addition, while methylation microarrays work well with formalin-fixated paraffin-embedded (FFPE) tissue, current protocols for nanopore sequencing depend on the availability of native or fresh-frozen tumour tissue. To support widespread implementation and analysis of archival tissues, protocols for robust nanopore sequencing from FFPE-derived DNA remain to be evaluated. Finally, with 31/66 (46.9%) MCF and 46/82 (56.1%) MC (excluding non-tumour control classes), not the full spectrum of brain tumours was studied in our cohort, and especially some rare entities were under-represented, and further studies are needed. Given the low prevalence of many tumour types, the growing number of newly described tumour entities and the ongoing refinement of brain tumour classifiers, however, it will be challenging to determine class-specific sensitivity and specificity of methylation-based classification on any technology platform in a statistically sound manner for very rare entities.

Cost implications for nanopore methylation-based classification

Competitive costs per assay are essential for the clinical adoption of a diagnostic test. Based on current list prices, the per-sample cost for nanopore singleplex sequencing is ∼ €240. This includes the (reusable) flow cell for use on a MinION device and sequencing chemistries. The initial hardware investment is ∼€1000 for the sequencing device and ∼€2000 for an analysis workstation (comprising a PC with consumer graphics processing unit [GPU] for hardware acceleration) in cases where a high-performance computing environment is not available. Given the results of sequencing time and required CpG sites, further optimisation of singleplex sequencing (reducing sequencing time per sample, increasing sample numbers per flow cell) may be feasible and could decrease cost per assay even further. Cost reduction by sample multiplexing is interesting; however, we observed a substantial reduction in read yield in the multiplex setting due to barcoding efficiency and probably the negative impact of low DNA contaminants in a single sample on the entire run. The above-stated per-sample cost only includes the consumables needed to perform low-pass whole-genome nanopore sequencing and neglects the further costs such as hardware investments for computing and the required workforce for library preparation.

CONCLUSION

In conclusion, we see great potential for routine implementation of nanopore sequencing in DNA methylation-based classification in brain tumour diagnostics not only to shorten the time to diagnosis but to augment neuropathological decision making and improve diagnostic precision. Further prospective evaluation in the context of a multicentric trial is warranted and ongoing. The approach might be particularly attractive to laboratories that see only a few neuro-oncological cases per week. Software compatibility with GPU-equipped multicore PCs significantly reduces the cost for the compute infrastructure and eliminates the need for access to a high-performance computing architecture.

ACKNOWLEDGEMENTS

We thank Aydah Sabah and Daniel Teichmann for expert technical assistance. Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health. N.B. and P.E. are participants in the BIH-Charité Clinical Scientist Program funded by the Charité – Universitätsmedizin Berlin and BIH. The project is funded by the Brain Tumour Charity, UK. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

DC declared a patent for a method to classify tumours according to DNA methylation signatures. All other authors declare no conflicts of interest.

ETHICS STATEMENT

The generation of prospective data within this study has been approved by the ethics committee at Charité – Universitätsmedizin Berlin (EA2/041/18).

AUTHOR CONTRIBUTIONS

LPK and PE conceived the study. LPK wrote the initial manuscript. LPK, NB, MM, JO, MC, AI and PE recruited study participants and collected clinical data or tumour material. LPK, IBH, EG, MB and JMP performed laboratory experiments and acquired sequencing data. LPK, JH and PE analysed sequencing data. LPK, PK, NI, DY, SL and PE performed software development. JH, SF, EGH and DC performed neuropathological review. CH, FH, DC and PE supervised data analysis and interpretation. All authors reviewed and approved the final manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/nan.12856.

DATA AVAILABILITY STATEMENT

The current nanoDx classification and analysis pipeline is publicly available at https://gitlab.com/pesk/nanoDx (version v.0.4.0rc1 was used for pre-processing of all sequencing data). Source code for the outlined RF implementation and to reproduce all analyses and figures in this manuscript is available at https://gitlab.com/pesk/nanoBenchmark. Raw sequencing data from 56 samples have been deposited at the European Genome-Phenome Archive (EGAS00001006540 and EGAS00001002213). Methylation microarray raw data and methylation calls were deposited at Gene Expression Omnibus (GSE209865).

ORCID

Luis P. Kuschel https://orcid.org/0000-0001-7794-2205
Philipp Euskirchen https://orcid.org/0000-0002-9138-805X

REFERENCES

1. Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555(7697):469-474. doi:10.1038/nature26000
2. Koelsche C, Schrimpf D, Stichel D, et al. Sarcoma classification by DNA methylation profiling. *Nat Commun*. 2021;12(1):1-10.
3. Moran S, Martínez-Cardús A, Sayols S, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol*. 2016;17(10):1386-1395. doi:10.1016/S1470-2045(16)30297-2
4. Jaunmuktane Z, Capper D, Jones DTW, et al. Methylation array profiling of adult brain tumours: diagnostic outcomes in a large, single centre. *Acta Neuropathol Commun*. 2019;7(1):24 doi:10.1186/s40478-019-0668-8
5. Euskirchen P, Bielle F, Labreche K, et al. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathol*. 2017;134(5):691-703. doi:10.1007/s00401-017-1743-5
6. Rand AC, Jain M, Eizenga JM, et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods*. 2017;14(4):411-413. doi:10.1038/nmeth.4189

7. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14(4):407-410. doi:10.1038/nmeth.4184
8. Schulze Heuling E, Knab F, Radke J, et al. Prognostic relevance of tumor purity and interaction with MGMT methylation in glioblastoma. *Mol Cancer Res*. 2017;15(5):532-540. doi:10.1158/1541-7786.MCR-16-0322
9. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol*. 2016;131(6):803-820. doi:10.1007/s00401-016-1545-1
10. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191
11. Scheinin I, Sie D, Bengtsson H, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res*. 2014;24(12):2022-2032. doi:10.1101/gr.175141.114
12. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36(4):338-345. doi:10.1038/nbt.4060
13. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/bioinformatics/bts480
14. Pedregosa F. Scikit-learn: machine learning in python. *J Machine Learn Res*. 2011;12:2825-2830.
15. Maros ME, Capper D, Jones DTW, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat Protoc*. 2020;15(2):479-512. doi:10.1038/s41596-019-0251-6
16. Zadrozny B, Elkan C, eds. Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02; 2002 2002-01-01: ACM Press.
17. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kuschel LP, Hench J, Frank S, et al. Robust methylation-based classification of brain tumours using nanopore sequencing. *Neuropathol Appl Neurobiol*. 2023;49(1):e12856. doi:10.1111/nan.12856