DIANA: An integrated pipeline for analysis of long-read whole-genome sequencing data for molecular neuropathology

Christian Domilongo Bope[1,2,4], Henning Leske[3], Richard Mark Nagymihaly[1,2], Einar Osland Vik-Mo[1,2,5] & Skarphéðinn Halldórsson[1,2*]

[1]Vilhelm Magnus Laboratory, Institute for Surgical Research, Oslo University Hospital, Oslo, Norway
[2]Department of Neurosurgery, Oslo University Hospital, 0372 Oslo, Norway.
[3]Department of Pathology, Oslo University Hospital, 0372 Oslo, Norway.
[4]Department of Mathematics and Department of Computer Science, Faculty of Sciences, University of Kinshasa, B.P. 190 / Kinshasa XI, Kinshasa, Democratic Republic of Congo
[5]Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

* To whom correspondence should be addressed. Email: skahal@ous-hf.no

ABSTRACT

Summary: Central nervous system (CNS) tumor diagnosis requires comprehensive genomic profiling including DNA-methylation classification, copy-number variants (CNV), gene fusion analysis, small variant detection and *MGMT* promoter methylation status. Long-read sequencing platforms such as nanopore sequencing by Oxford Nanopore Technologies and SMRTseq by PacBio can capture all these in a single assay, but integrating diverse analytical tools to leverage the advantages of long-read sequencing remains complex. We present DIANA (Diagnostic Integrated Analytics of Neoplastic Alterations), a pipeline providing fully automated end-to-end processing of long-read whole-genome sequencing data from aligned sequence reads. DIANA produces a human readable report that combines methylation classification with prioritized genetic variants to support CNS tumor diagnostics and clinical decision-making.

Availability and implementation: DIANA is an open-source Nextflow pipeline, freely available through Docker or Apptainer/Singularity technologies. The source code, comprehensive documentation, and installation protocols are available on GitHub: https://github.com/VilhelmMagnusLab/DIANA.git .

Supplementary information: Supplementary data are available at Bioinformatics online.

INTRODUCTION

Accurate diagnosis of central nervous system (CNS) tumors requires comprehensive genetic analysis, including the detection of single nucleotide variants (SNVs), copy number variations (CNVs), gene fusions and epigenetic modifications (Louis, et al., 2021). Recently, DNA methylation profiling has emerged as a powerful tool and is now integral to the classification of some CNS tumors according to the World Health Organization (WHO) 2021 guidelines (Capper, et al., 2018). To fully capture all necessary alterations, current diagnostic workflows rely on separate platforms including next-generation sequencing, methylation arrays, pyro-sequencing, and fluorescent in situ hybridization (Bertero, et al., 2024). This creates significant logistical and financial burdens that can limit both their implementation and their routine use in clinical practice.

Long-read whole-genome sequencing captures DNA methylation, structural variants, small variants and copy number alterations simultaneously within a single assay (O'Neill, et al., 2024). Despite these advantages, integrating diverse analytical tools and interpretation of the results remains challenging.

To address these limitations, we developed DIANA, a Nextflow-based pipeline that provides end-to-end automation of long-read WGS data from CNS tumors. To our knowledge, this is the first publicly available end-to-end pipeline for tumor diagnostics using long-read WGS data. The pipeline integrates DNA methylation-based tumor classification, O^6-methylguanine-DNA methyltransferase (*MGMT*) promoter methylation quantification, and detection and annotation of structural variants, copy number variations, and single nucleotide variants into a unified workflow. Each analytical module produces structured raw result files covering all variant categories, methylation calls, and copy number profiles. These raw outputs are subsequently aggregated, annotated and filtered to generate a concise genetic report that highlights the key diagnostic findings most relevant to patient management. All data remain fully accessible, enabling clinicians and researchers to perform independent verification or custom downstream analyses as required. Full containerization ensures reproducibility and portability across diverse computing environments, from local workstations to high-performance computing (HPC) clusters.

DIANA workflow

DIANA pipeline is implemented in Nextflow (Langer, et al., 2025) (ver. 25.04.6) with DSL 2 enabled, integrating custom Python, R, and Bash scripts into a unified, modular workflow that enables scalable execution across local workstations and HPC clusters. The pipeline consists of four main modules that can be executed independently or sequentially as illustrated in Figure 1. The input consists of single or multiple BAM files generated by long-read sequencing from either Oxford Nanopore Technologies (ONT) or PacBio. All outputs are consolidated into a comprehensive PDF report generated with R language and R Markdown templates.

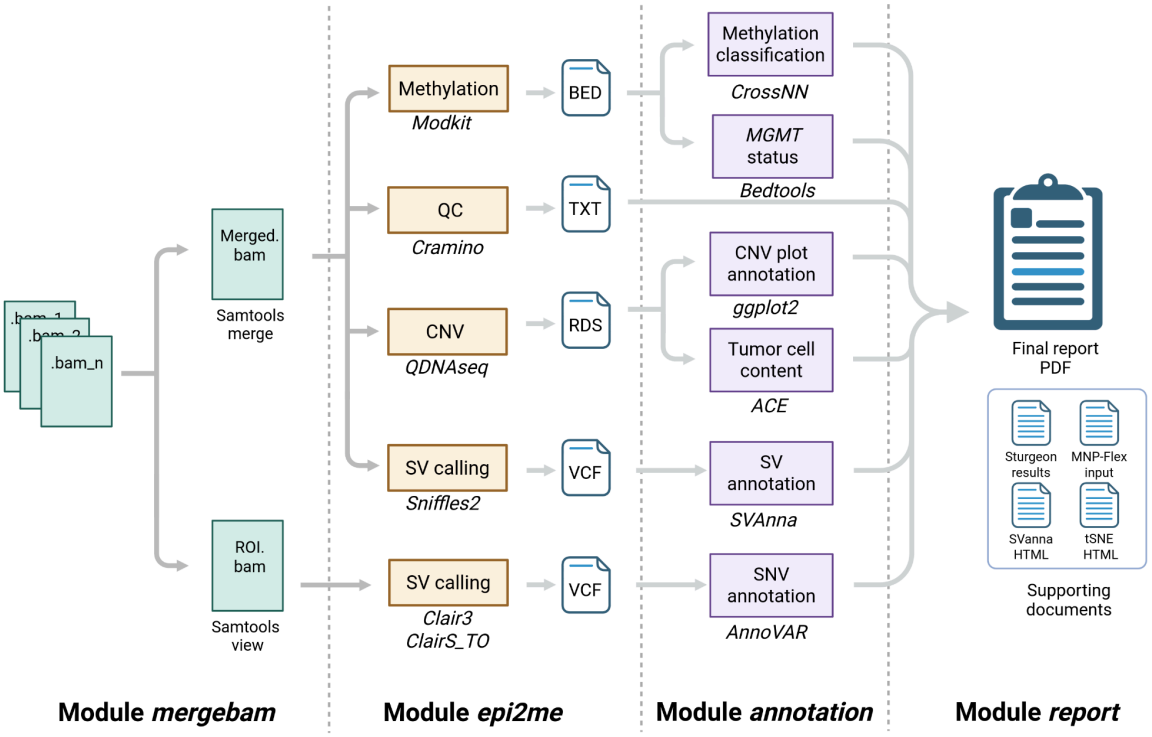

Figure 1. Overview of the DIANA workflow. The pipeline is organized into four distinct modules: *mergebam* for data preprocessing and region-of-interest (ROI) extraction; *epi2me* for primary genomic and epigenetic analysis, including methylation profiling (*Modkit*), structural variant calling (*Sniffles2*), and copy number variation (*QDNAseq*); *annotation* for biological interpretation and diagnostic relevance and *report* for the automated synthesis of results into a final PDF report and supporting documentation.

First module, *mergebam:* The MinKNOW software (ONT) can be configured to write BAM files at regular intervals throughout the run (e.g., hourly) rather than producing a single file upon sequencing completion. This is the recommended approach, as it protects against data loss in the event of an unexpected interruption before the run ends. To avoid any risk of sample mix-up, the *mergebam* module uses a combination of sample identifier and flow cell identifier to unambiguously assign each BAM file to the correct sample. All files that share the same sample and flow cell identifiers are merged into a single merged and indexed BAM using *samtools merge* and *samtools index* (Danecek, et al., 2021). A second output BAM is generated by extracting only the genomic regions of interest (ROI) defined in a user-supplied BED file of clinically relevant genomic regions for SNV calling.

Second module, *epi2me:* This module performs methylation tag extraction, structural variant calling, copy number analysis and small variant calling in parallel. Genome-wide CpG methylation quantification is performed using *modkit* (https://github.com/nanoporetech/modkit) on the whole-genome merged BAM file. Modkit interrogates the base modification tags embedded in the BAM file to produce a bedmethyl file per-CpG quantifying levels of 5-methylcytosine (*5mC*) and 5-hydroxymethylcytosine (*5hmC*) throughout the genome. Structural variant (SV) detection is carried out with *Sniffles2* (Smolka, et al., 2024) to identify insertions, deletions, inversions, and translocations and outputs a compressed VCF file. Genome-wide copy number profiling is performed using *QDNAseq* (Scheinin, et al., 2014), with a default binsize set to 50 kbp. *QDNAseq* applies GC-content and mappability correction to call copy number segments; the resulting segment BED file, bin-level BED file, and RDS object are stored as the primary raw input for the downstream analysis. Small variant calling (SNV and indel) is performed on the targeted ROI BAM file using *Clair3* (Zheng, et al., 2022) for germline variants and C*lairS*-TO (Chen, et al., 2025) for somatic variant detection. This generates compressed VCF files for downstream analysis. Sequencing quality metrics (read quality, coverage, read-length distribution) are computed using *cramino* (De Coster and Rademakers, 2023).

Third module, *annotation:* This module takes the outputs of the *epi2me* module and performs a set of deeper interpretation and annotation steps. The following analysis is performed: (1) *MGMT* promoter methylation is quantified by extracting CpG sites overlapping the CpG island in the *MGMT* promoter, computing the mean methylation and providing a methylated versus unmethylated status (Halldorsson, et al., 2024). Results are saved as a CSV file. (2) Methylation-based classification is

performed by two independent neural network classifiers: *CrossNN* (Yuan, et al., 2025) and *Sturgeon* (Vermeulen, et al., 2023). By default, both classifiers use the reference data provided by Capper *et al.* (Capper, et al., 2018) to assign a 2021 WHO tumour class with probability scores. In addition, *CrossNN* can be set to use the Pan-cancer reference data for broad tumor type classification (using the *--pancan* flag). A *UMAP* plot is also generated placing the sample context with the reference data. (3) Structural variant annotation is carried out with *Svanna* (Danis, et al., 2022), which cross-references the *Sniffles2* VCF against Human Phenotype Ontology (HPO) based gene prioritisation to highlight clinically relevant rearrangements. Gene fusion events with exon-level breakpoint coordinates are extracted and specifically annotated. (4) An annotated CNV plot is generated from the *QDNAseq* segment and bin outputs which places the corrected *log2ratio* of each bin and segment onto chromosomal coordinates using ggplot2 in R. Genes or gene-pairs of interest are annotated according to a user provided BED file. Chromosomal segments with gains or losses are highlighted. (5) The RDS file generated by *QDNAseq* is used as input for tumour cell content estimation using *ACE* (Poell, et al., 2019). (6) The small variant VCF files generated by *Clair3* and *ClairS_TO* are merged, filtered, and annotated against *ClinVar* (Landrum, et al., 2025) and *COSMIC* databases (Tate, et al., 2019) to produce a single ranked variant table. Integrated Genomics Viewer (*IGV)* (Thorvaldsdottir, et al., 2013) coverage snapshots are generated for key diagnostic loci (*EGFR*, *IDH1*, *IDH2*, *TERT* promoter).

Fourth module, *report generation***:** All outputs from the *analysis* module are consolidated into a comprehensive PDF report generated via an R Markdown template. The report is structured in two complementary sections. An Executive Summary provides a rapid clinical overview on a single page, opening with the mean sequencing coverage and the estimated tumour cell content, followed by: the methylation-based tumour *CrossNN* classification at class and class-family level with confidence scores, the genome-wide copy number profile, a filtered SNV table restricted to diagnostically relevant target genes and pathogenic *ClinVar* variants, a summary of detected fusion events, and the *MGMT* promoter methylation status. An Extended Report follows that provides full analytical detail for each domain: detailed read statistics (number of alignments, percentage of mapped reads, total read count, and mean coverage), the full tumour content estimate; *CrossNN* classification Methylation Class and Methylation Class Family bar plots for the top five classification scores; a *UMAP* dimensionality reduction plot for visual inspection of classifier positioning; all chromosome arm-level CNV alterations; a

high-confidence CNV table filtered for amplifications and homozygous deletions; focused chromosome 7 (*EGFR*) and chromosome 9 (*CDKN2A/B*) profiles; overage snapshots for *IDH1* p.R132, *IDH2* p.R172, and *TERT* promoter hotspots; the full per-CpG *MGMT* methylation table (CpG sites 1–98 and 76–79); and the complete annotated SNV and structural variant tables. Alongside the main report PDF, a set of supplementary outputs is published: a *Sturgeon* methylation classification PDF, an interactive *Svanna* Structure Variant annotation HTML, an interactive *CrossNN* classification *UMAP* HTML plot, and a formatted methylation BED file for compatibility with *MNP-Flex* classification platforms.

DIANA run-modes

DIANA supports flexible execution through multiple run modes. (i) *Smart sample monitor*: an automated monitoring utility that continuously watches sample directories for the presence of a sequencing completion marker file (*final_summary_*_*_*.txt*) that is written by MinKNOW when the sequencing run reaches its predefined endpoint (time or data limit reached) or is manually stopped. Upon detection, it automatically queues and launches the full pipeline from *mergebam* to R Markdown *report* for each completed sample. (ii) *–run_mode_mergebam*: runs only the BAM merging and ROI extraction step, for example when BAM preparation must be performed separately. (iii) *–run_mode_order*: triggers complete end-to-end execution of all modules in sequence (*mergebam epi2me annotation report*), processing samples from unmerged BAM files through to the final report. (iv) *–run_mode_epianalyse:* runs the *epi2me* and *annotation* modules sequentially, skipping *mergebam*; intended for cases where BAM files have already been merged. (v) Individual sub-module execution: each *epi2me* analysis type and each *annotation* step can be run independently using dedicated flags, for example *–run_mode_epi2me modkit* to rerun methylation calling only, *–run_mode_analysis mgmt* for *MGMT* quantification, and *–run_mode_analysis rmd* to regenerate the PDF report without rerunning the upstream analyses. An example report generated from a glioblastoma genome with nanopore whole-genome sequencing is found within the **Supplementary Document**.

DISCUSSION

DIANA provides a reproducible and standardized framework that ensures that all diagnostically relevant dimensions of CNS tumor characterization including DNA methylation classification, *MGMT* methylation status of the promoter region, copy

number alterations, structural variants and single nucleotide variants are systematically captured, annotated, and presented. The open-source, containerised design ensures that results are fully reproducible across computing environments. The modular architecture further allows individual analytical steps to be rerun in isolation when needed, without reprocessing the entire dataset, reducing both compute time and resource consumption. The complete pipeline modules (*mergebam*, *epi2me*, and *annotation*) for a sample with 28.78× coverage were completed in approximately 82 minutes of wall-clock time on a single workstation using parallel processing (Supplementary Fig. S1). Resource usage varied substantially across processes: the rate-limiting steps were BAM merging (19 minutes, peak RAM 1.3 GB) in the *mergebam* module and somatic SNV calling with *ClairS-TO* (59 minutes, peak RAM 28.6 GB) in the *epi2me* module. The annotation module was notably lightweight, completing in under 4 minutes for all processes. The longest individual process was the generation of *UMAP* methylation classifier visualization (1 minute 40 seconds, peak RAM 37.9 GB). Based on these benchmarks, the minimum recommended system configuration to run the full DIANA pipeline is 32 GB of RAM, 8 CPU cores, and approximately 500 GB of available disk space for intermediate files. All processes completed successfully with no out-of-memory failures, and the annotation module ran sequentially after *epi2me*, contributing less than 5% of total runtime.

Accessibility was a central design consideration: DIANA runs on standard local workstations as well as HPC clusters, requires no prior bioinformatics expertise to operate, and is freely available under MIT open licence with fully automated setup scripts.

ACKNOWLEDGEMENTS

We thank Dr. Philipp Euskirchen and Dr. Areeba Patel for insightful discussions on methylation classifiers and analytic software, and Dr. Marius Lund Iversen for help with variant prioritization. The authors declare the use of Claude (Anthropic, Claude Sonnet 4.6) as an AI-assisted tool for code debugging during the development of the DIANA pipeline. All AI-generated contributions were reviewed and validated by the authors.

FUNDING

This work was supported by grants from Norwegian South-Eastern regional health authorities [grant numbers 2021039, 2023059], The Norwegian Cancer Society [grant number 323453] and Marthe Fondet.

Conflict of Interest: HL, EOVM and SH received travel reimbursements from Oxford Nanopore Technologies to present at conferences in 2023-2025.

REFERENCES

Bertero, L., *et al.* Molecular neuropathology: an essential and evolving toolbox for the diagnosis and clinical management of central nervous system tumors. In, *Virchows Archiv*. 2024.

Capper, D., *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* 2018;555(7697).

Chen, L., *et al.* ClairS-TO: a deep-learning method for long-read tumor-only somatic small variant calling. *Nat Commun* 2025;16(1):9630.

Danecek, P., *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10(2).

Danis, D., *et al.* SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing. *Genome Medicine* 2022;14(1):1–13.

De Coster, W. and Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 2023;39(5).

Halldorsson, S., *et al.* Accurate and comprehensive evaluation of O6-methylguanine-DNA methyltransferase promoter methylation by nanopore sequencing. *Neuropathology and applied neurobiology* 2024;50(3).

Landrum, M.J., *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res* 2025;53(D1):D1313–D1321.

Langer, B.E., *et al.* Empowering bioinformatics communities with Nextflow and nf-core. *Genome Biol* 2025;26(1):228.

Louis, D.N., *et al.* The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology* 2021;23(8).

O'Neill, K., *et al.* Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes. *Cell genomics* 2024;4(11):100674–100674.

Poell, J.B., *et al.* ACE: Absolute copy number estimation from low-coverage whole-genome sequencing data. *Bioinformatics* 2019;35(16).

Scheinin, I., *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 research* 2014;24(12):2022–2032.

Smolka, M., *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology* 2024;42(10):1571–1580.

Tate, J.G., *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* 2019;47(D1):D941–D947.

Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14(2):178–192.

Vermeulen, C., *et al.* Ultra-fast deep-learned CNS tumour classification during surgery. *Nature* 2023;622(7984):842–849.

Yuan, D., *et al.* crossNN is an explainable framework for cross-platform DNA methylation-based classification of tumors. *Nature Cancer* 2025;11:1–12.

Zheng, Z., *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci* 2022;2(12):797–803.