biomodal

TECHNOLOGY NOTE

Unlocking 6-base insights through DMR calling with modality XPLR

modality XPLR is an accessible and scalable software tool for analysing biomodal epigenetic data, from duet evoC (6-base readout with independent counts of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC)) and duet +modC (combined 5mC and 5hmC counts for a 5-base modC readout).

Differentially Methylated Regions (DMRs) are genomic regions where DNA methylation levels in modC, 5mC or 5hmC, differ significantly between two or more groups—such as healthy vs disease, or treated vs untreated samples.

With modality XPLR, users can call DMRs for modC, or independently for 5mC and 5hmC, enabling a nuanced view of epigenetic changes. Here, we demonstrate how the modality XPLR DMR calling and plotting functions allow you to derive actionable insights from biomodal methylation data.

DMRs are important for Epigenetic Discovery

DMRs are used to understand how DNA methylation changes influence biological processes and disease states. By identifying regions of the genome where methylation levels differ significantly between sample groups—such as healthy versus diseased tissue—DMRs reveal patterns of epigenetic change that could inform gene regulation, cellular reprogramming, and disease progression. They serve as powerful biomarkers for diagnostics, prognostics, and therapeutic response, especially in contexts like cancer, where methylation dynamics can signal early transformation or treatment efficacy.

How DMR Calling Works: Statistical Overview

The DMR calling workflow in modality XPLR is designed for flexibility and scientific rigour:

- **Start with Two Groups:** Define your experimental groups (e.g., healthy vs. disease, early vs. late stage).
- **Define Regions:** Segment the genome into regions using either custom definitions (e.g. gene bodies, promoters) or by specifying window size and/or pre-segmentation options for the whole-genome (e.g. fixed 1Kb windows, or dynamic windows based on CpG distances).
- **Aggregate Modification Counts:** For each region, counts of modified and unmodified cytosines at CpG sites are aggregated across samples.
- **Statistical Testing:** A logistic regression model compares methylation proportions between groups, optionally adjusting for covariates, yielding a p-value for each region.
- **Multiple Testing Correction:** p-values are adjusted to produce q-values, controlling the false discovery rate (FDR).
- **Overdispersion Correction:** An optional setting to improve statistical reliability by accounting for variability beyond the expected noise. This reduces the FDR, with some trade-off in sensitivity to reveal high-confidence DMRs.

This workflow ensures that DMR detection in modality XPLR is both statistically robust and adaptable to diverse experimental designs, enabling reliable identification of methylation differences across genomic regions.

Discover DMRs Using Broad CpG-Informed or Annotation-Based Approaches

When calling DMRs, the regions can be defined as:

- Fixed tiling across the whole genome or a specific contig, such as 1kb windows.

- Custom regions, such as a list of promoters, enhancers, gene bodies or CpG islands.
- Dynamic, genome pre-segmentation by distribution of CpGs according to minimum CpG distance, maximum region size and minimum number of contexts settings.

Figure 1 shows a volcano plot for 5hmC DMRs across hg38 gene bodies, showing 5hmC hypermethylation in cerebellum tissue when compared to lung tissue. In a volcano plot, the modification difference of each region is plotted against the statistical significance of that change. Values in the top left and top right represent the largest and most significant losses or gains in methylation between groups, respectively.

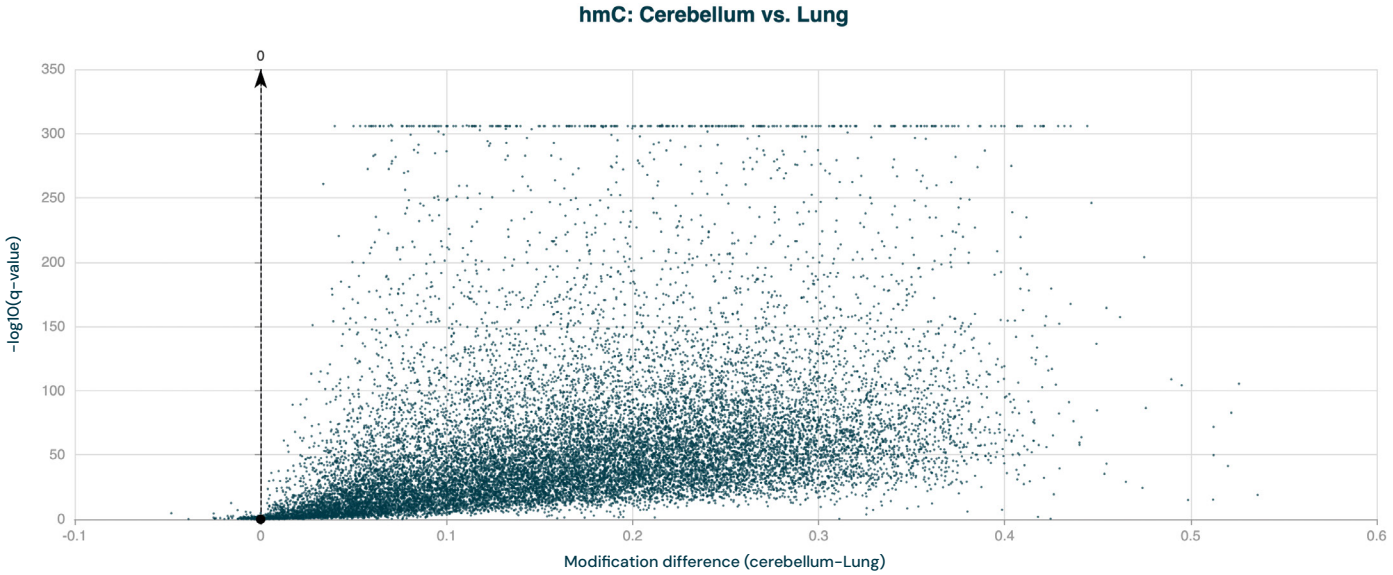

Figure 1: Volcano plot showing 5hmC DMRs for gene bodies between Lung (control) and Cerebellum (test) tissues.

Use Biological Priors to Accelerate DMR Discovery

Genome-wide DMR screening using fixed windows increases the number of statistical comparisons and the risk of false discoveries. Stringent corrections for FDR (e.g. overdispersion correction) may be required, which can reduce statistical power to detect true DMRs. Given what is known about the biology of a research question, it is often possible to use a biological prior to help design an analysis, focussing on regions or genomic features known to have an important biological role. Testing over a smaller set of regions lowers the multiple testing burden, and can increase the likelihood of discovering a meaningful and biologically relevant methylation change. modality XPLR supports region definition in BED3+ format, allowing you to target custom regions that may be implicated in your area of study.

The next example shows how modC DMRs between healthy and late-stage CRC tissue [1], were used as biological priors to discover 5mC DMRs in cfDNA, from healthy and affected individuals. Figure 2 shows a volcano plot with a high density of 5mC hypomethylation between Healthy Controls and Stage IV CRC in cfDNA, confirming the hypothesis that cfDNA 5mC DMRs overlap with tissue modC DMRs.

Compare Early and Late-Stage Disease DMRs

Analysis outputs of modality XPLR follow a standardised BED3+ format to link statistical data to genomic regions. Therefore, a DMR result file from one analysis can be directly used to test your next hypothesis, whether by further DMR analysis or extraction of feature statistics for downstream modelling.

After identifying DMRs between healthy and late-stage-disease conditions, DMR calling can be applied to identify early-stage biomarkers that signal disease-associated epigenetic change. Given the relationship between 5mC and 5hmC, we can use late-stage loss of 5mC to guide where we expect to see 5hmC changes at earlier disease stages. In this example late-stage 5mC DMRs (Figure 2) were filtered for those showing

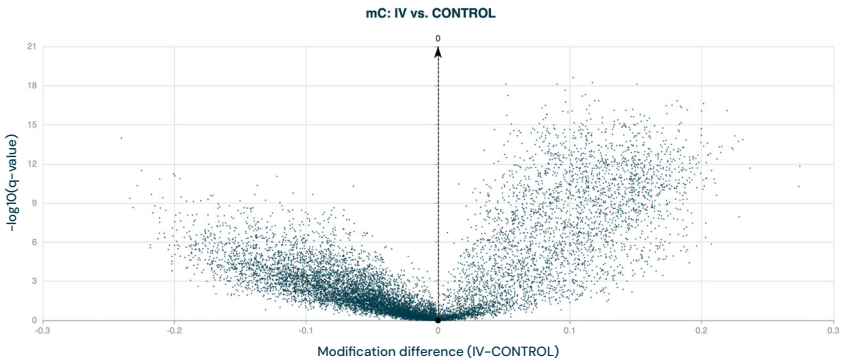

Figure 2: Volcano plot showing 5mC DMRs for Healthy Control and Stage IV CRC cfDNA samples, using TCGA tissue-derived DMRs as biological priors.

hypomethylation only, and used as input for 5hmC DMR calling at earlier stage (Stage II). The results of this analysis are represented in the volcano plot in Figure 3, validating that late stage 5mC loss is foreshadowed by earlier 5hmC gain, unlocking a novel biomarker of early biological change.

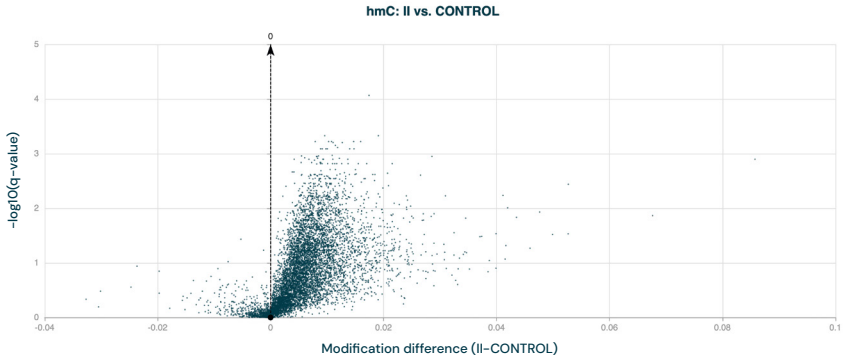

Figure 3: Volcano plot showing 5hmC DMRs between Healthy Controls and Stage II CRC cfDNA, for regions identified as hypomethylated for 5mC in Stage IV CRC cfDNA.

View DMRs in Genomic Context with Track Plots

Track plots in modality XPLR allow the visualisation of 5mC and 5hmC DMRs in genomic context (Figure 4). By aligning DMRs with gene annotations and base-resolution modification traces, researchers can explore how epigenetic changes intersect with promoters, enhancers, gene bodies, and other regulatory elements. This spatial perspective transforms statistical results into biologically meaningful insights—revealing whether a DMR overlaps a functional region, correlates with gene regulation, or reflects disease-associated epigenetic reprogramming. Track plots bridge the gap between discovery and interpretation, enabling users to assess the distribution, direction, and magnitude of methylation changes with confidence.

For example, a DMR analysis targeting promoter regions (defined as 1,000 bp upstream of transcription start sites) identified the *ENGASE* promoter as significantly hypomethylated for 5mC in late-stage CRC cfDNA compared to Healthy Controls. ENGases (endo-β-N-acetylglucosaminidases) are enzymes implicated in cancer biology due to their role in modifying N-glycans—structures frequently altered in tumour progression. The track plot in Figure 4 highlights this promoter DMR within the broader gene body, illustrating the local methylation landscape and supporting its potential as a diagnostic or prognostic biomarker.

Methylation fraction by genomic position (points) and smoothed (line) at CpG contexts for Control (dark teal) and Test (light teal) groups.

Methylation difference by genomic position (points) and smoothed (line) at CpG contexts between the two groups. Positive values indicate hypermethylation and negative values indicate hypomethylation relative to the Control group.

Magnitude of modification difference for defined DMR regions (block), Positive values indicate hypermethylation and negative values indicate hypomethylation relative to the Control group. Significance level indicated by the number of asterisks.

Annotation of genomic features via user-supplied GTF/GFF reference, showing genes and exons (blue boxes).

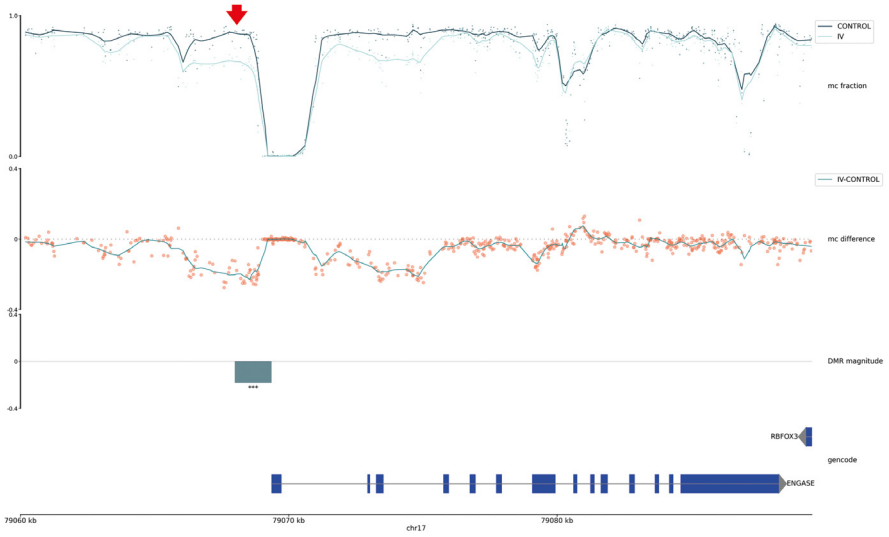

Figure 4: Track plot for the *ENGASE* gene, showing a hypomethylated 5mC DMR between Healthy Control and Stage IV CRC cfDNA. The four tracks are methylation fraction, methylation difference, DMR magnitude and genome annotation. The DMR significance (p-value) is indicated by the number of asterisks ($p \leq 0.001$), the difference between 5mC fractions is highlighted by divergence of the mean methylation between the two groups (red arrow), immediately prior to the *ENGASE* TSS (annotation track).

Evaluating DMR Calling Accuracy with Semi-Simulated Data

To assess the accuracy of DMR calling in modality XPLR, a semi-simulated dataset was created by embedding synthetic 5mC DMR regions of varying length and effect size into real cfDNA samples from Healthy Control individuals (Test group). The Test group was analysed against different, unmodified Healthy Control samples for DMRs using modality XPLR, across different window size settings, with overdispersion correction enabled to account for biological variability. Sensitivity and false discovery rate (FDR) were calculated for DMR window size, by mean strand-merged CpG coverage (Figure 5A), and number of samples (Figure 5B). Results demonstrated that FDR is controlled with smaller windows, however sensitivity increases with the number for contexts per region (larger windows). CpG coverage >20x provides diminishing returns in performance, however increasing the number of samples powers discovery.

We then evaluated DMR accuracy with increasing effect size, using regions defined by modality XPLR's pre-segmentation algorithm, requiring ≥5 contexts per region, a maximum region size of 5000 bp, and ≤500 bp between CpGs—yielding an average region size of ~2.5 Kb on the hg38 genome. With overdispersion correction enabled, sensitivity improved and FDR declined with increasing effect size (Figure 6). These results demonstrate that modality XPLR delivers robust sensitivity and controlled FDR across diverse sequencing depths, sample sizes, and region configurations, validating its reliability for epigenetic biomarker discovery in diverse experimental conditions.

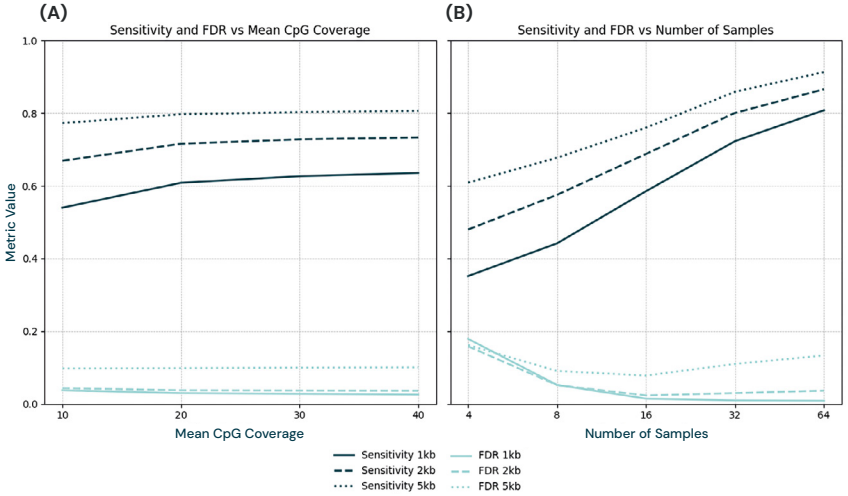

Figure 5: modality XPLR 5mC DMR calling sensitivity and FDR, for region sizes of 1Kb, 2Kb, and 5Kb, by **(A)** mean strand-merged CpG coverage with 8 samples per group and **(B)** sample size with up to 32 samples per group (64 total) and a mean strand-merged CpG coverage of 21x.

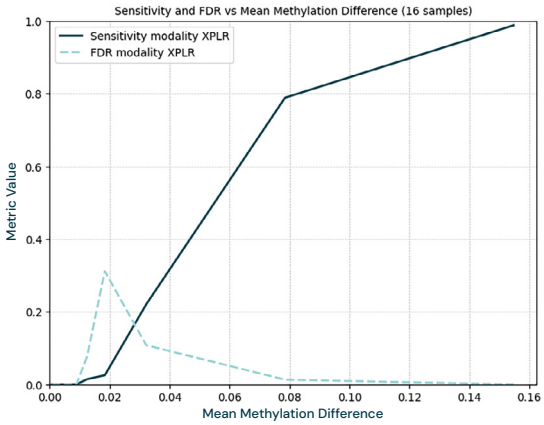

Figure 6: Sensitivity and FDR for 5mC DMR calling for pre-segmented regions by mean methylation difference (effect size).

To assess real-world usability, we tested modality XPLR and methlyKit (a community tool) [2] using an 8-sample pilot dataset on a standard laptop. Using modality XPLR, DMR calling over 19,382 promoter regions completed in under 7 minutes. However, methylKit failed to load the input files due to memory constraints.

This comparison highlights modality XPLR's scalability and efficiency, making it a practical solution for large-scale epigenomic studies.

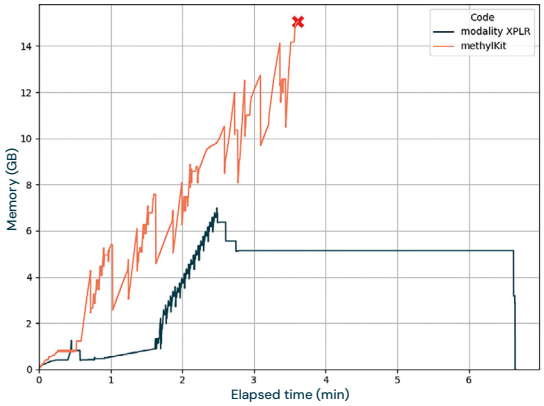

Figure 7: Memory usage over time for genome-wide DMR calling on 19,382 promoters with 8 samples on a standard laptop (4 cores, 16 GB RAM). Whilst modality XPLR efficiently completes the analysis in a round 7 minutes, methylKit cannot complete the operation due to memory exhaustion.

Conclusion

modality XPLR offers a robust and flexible platform for the discovery and interpretation of differentially methylated regions (DMRs) across both 5mC and 5hmC marks, enabling researchers to gain deeper insights into epigenetic changes associated with disease states and treatment responses. By supporting a range of region definitions, integrating biological priors, and providing advanced statistical controls, modality XPLR empowers users to identify meaningful methylation patterns, visualise results in biological context, and accelerate biomarker discovery for early detection and monitoring. The standardised outputs further facilitate downstream analysis and integration into broader research workflows, making modality XPLR a valuable tool for advancing epigenetic research and translational research applications.

References

1. Puddu, F., Johansson, A., Modat, A., Scotcher, J., Sethi, R., Yu, S., Harding, N., Hill, M., Lleshi, E., Lumby, C., Teyssandier, J., Wilson, M., Crawford, R., Charlesworth, T., Osborne, R.J., Balasubramanian, S., Creed, P. (2024). 5-methylcytosine and 5-hydroxymethylcytosine are synergistic biomarkers for early detection of colorectal cancer [Preprint]. *bioRxiv*, doi:10.1101/2024.10.30.621123.

2. Akalin, A., Kormaksson, M., Li, S., Wabo, A., Bierling, A., Blume, A., Wreczycka, K. (2025). methylKit: DNA methylation analysis from high-throughput bisulfite sequencing [Software]. *GitHub*. https://github.com/al2na/methylKit

Disclaimer

modality XPLR is for research use only.

Chesterford Research Park
Cambridge, UK
CB10 1XL

+44 (0) 1223 800 700
biomodal.com
info@biomodal.com

biomodal