A High-Performance Toolkit for Large-Scale Analysis of 5- and 6-Base Genomes

Jean Teyssandier, Nicholas Harding, Sabri Jamal, Michael J. Wilson, Gary Frewin, Nicola Wong, William Stark, Mark S. Hill, Páidí Creed[1]

1 biomodal Ltd, The Trinity Building, Chesterford Research Park, Cambridge, UK

biomodal

1. Introduction

We present a computational toolkit to analyse 5mC and 5hmC modifications at scale and describe its performance on a novel liquid biopsy dataset. Methylation data has diverse applications in cancer, including early-stage diagnosis through liquid biopsy, classification to guide treatment pathways, and prognosis. However, analyzing methylation data poses significant challenges, as it is constrained by scalability and usability issues.

Our recently introduced technology, duet multiomics solution evoC, enables the reading of 6-base information (A, T, G, C, 5-mC and 5-hmC) from DNA, further amplifying the complexity and scale of datasets generated in a single sequencing experiment. To address this, we present a fast and scalable array-based python package (called modality) for the analysis of 6-base genomes, using multi-core out-of-memory processing to enable extremely efficient computation, even for datasets that are too large to fit into memory. Unlike existing tools that exceed typical laptop memory with ~10 samples, the modality package can efficiently analyse (e.g. DMR calling) a colorectal cancer liquid biopsy dataset of over 100 samples in minutes on a standard laptop.

modality combines efficient computation with tools for exploratory (e.g., plotting, methylation summaries) and downstream analyses (e.g., DMR identification, PCA). Designed for efficiency and ease of use, it enables users to rapidly transition from raw data to actionable insights and publication-ready results. As multiomic data become the standard in cancer research, our data structure supports the integration of additional data types, allowing us to handle combined genomic and epigenomic data from solutions like duet evoC. This will enable streamlined and efficient multiomic analysis to uncover deeper biological insights.

2. duet evoC 6-base sequencing [A,C,T,G,5mC,5hmC]

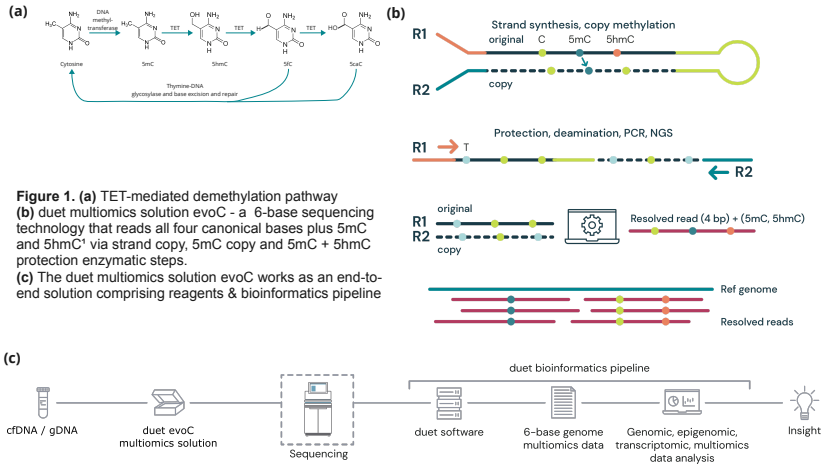
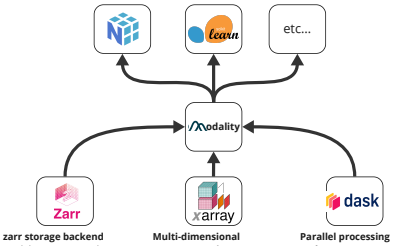


Figure 1. (a) TET-mediated demethylation pathway
(b) duet multiomics solution evoC - a 6-base sequencing technology that reads all four canonical bases plus 5mC and 5hmC[1] via strand copy, 5mC copy and 5mC + 5hmC protection enzymatic steps.
(c) The duet multiomics solution evoC works as an end-to-end solution comprising reagents & bioinformatics pipeline

3. Code description and features

modality is built around three core python packages: *zarr*, *xarray*, and *dask*. These are powerful modern data science packages which collectively allow *modality* to deal with larger-than-memory data arrays in an expressive and paralelised fashion. This foundation means that analyses that would previously require long run times and extensive compute infrastructure can now run quickly on one's laptop - speeding up iterative data analysis.

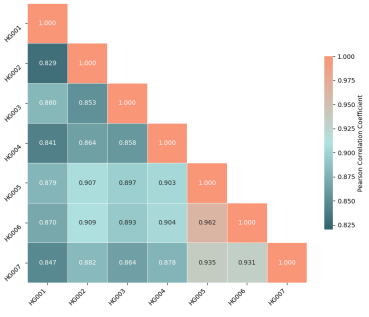
The core data structure used by *modality* is the ContigDataset. This contains arrays that represent methylation counts as well as accompanying arrays which encode the coordinates. This object also provides a set of easy-to-use and efficient methods for working with them. Each of the arrays are chunked *Dask* arrays allowing extremely efficient computation.



zarr storage backend
modality interacts with *zarr*, leveraging chunked and compressed sets of n-dimensional arrays, allowing analyses to efficiently scale to many samples (>100) even with very limited RAM.

Multi-dimensional Data Interaction
modality enables seemless interaction with datasets across multiple dimensions using *xarray*.

Parallel processing and memory efficiency *modality* leverages the parallel processing capabilities of *dask* for efficient computation.



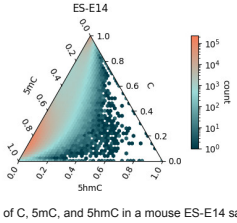
Pearson correlation coefficients for the 7 samples in the GIAB data

modality offers a large choice of analysis tools:

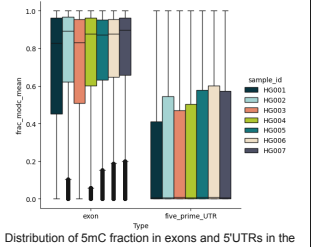
- Biological QC and exploratory analysis (Pearson correlation between samples, PCA...)
- Feature extraction: summary of methylation data over genomic ranges
- DMR analysis

modality can be used either as:

- a set of CLI commands
- a python package to be imported in scripts/notebooks

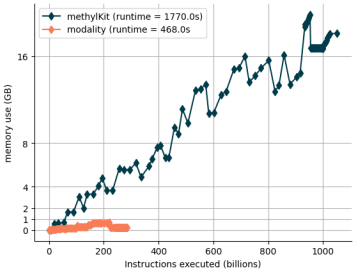


Ternary plot of C, 5mC, and 5hmC in a mouse ES-E14 sample



Distribution of 5mC fraction in exons and 5'UTRs in the GIAB data using modality's feature extraction capabilities

4. Performances



Benchmarking of modality vs. MethylKit on DMR calling

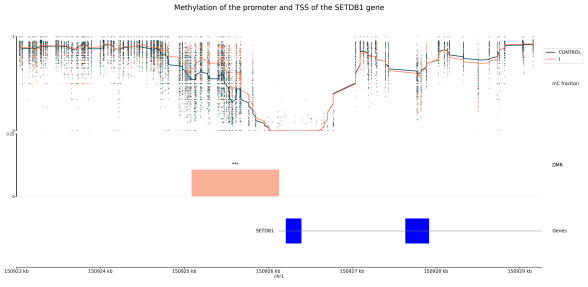
We conducted a performance comparison between the DMR caller in modality and the one from MethylKit. A key limitation of MethylKit is its high memory consumption when reading text files, which eventually leads to it exhausting the available system memory. In contrast, modality is memory-efficient, allowing for genome-wide DMR calling on a **standard laptop**.

To evaluate real-world usability, we tested both tools on a colorectal cancer (CRC) dataset consisting of **58 samples** divided into two conditions (control vs stage I). Using modality, we were able to perform genome-wide DMR calling in just **12 minutes** when tiling the genome into **2kb windows**. In comparison, MethylKit was unable to process the dataset in a single run, as it ran out of memory while attempting to read the input files. This highlights modality's scalability and efficiency, making it a more practical choice for large-scale epigenomic analyses.

5. Application to a CRC dataset

modality can process large cohorts of samples genome-wide on a standard laptop. Here we show an example of an analysis on a colorectal cancer data.

- We call DMRs between Control and Stage I patients on promoters (defined as 1kb upstream of TSS) of protein-coding genes from the GENCODE hg38 annotation database.
- We identify significant DMRs by filtering by q-value and mean methylation difference between the two groups.
- Below we show an example of a differentially methylated promoter upstream of the SETDB1 gene, which is known to promote CRC progression [2], using modality's genomic tracks plot feature.



6. Conclusion

To address the difficulties of analysing methylation data, we present *modality*, an efficient and scalable analysis package for 5- and 6-base genomes.

- The package is built on a core set of performant data science libraries and roots the user into a powerful ecosystem for data analysis in Python.
- *modality* has an intuitive API providing powerful analysis and visualisation methods, enabling users to gain insight from duet multiomics solution evoC directly on a laptop.
- Moving forward, the underlying data structure used by *modality* is very extensible, allowing other data modalities to be incorporated and analysed alongside the methylation data modalities shown here.

7. References

1. Füllgrabe J. et al. Simultaneous sequencing of genetic and epigenetic bases in DNA. Nat Biotechnol. 2023 Oct;41(10):1457-1464.
2. Cao, N., et al. SETDB1 promotes the progression of colorectal cancer via epigenetically silencing p21 expression. Cell Death Dis 11, 351 (2020).

biomodal