Simultaneous identification of methylation and somatic variants can improve sensitivity for cancer detection and monitoring

Aurelie Modat[1], Cillian Nolan[1], Ermira Lleshi[1], Robert Blanshard[1], Sascha Seidel[1], Fabio Puddu[1], Annelie Johansson[1], Ermira Lleshi[1], Robert Crawford[1], Tom Charlesworth[1], **Robert J Osborne**[1]

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

1. Introduction

5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), are known as the fifth and sixth DNA bases of the genome. 5mC and 5hmC have different distributions in different tissues and can act as tissue-specific fingerprints. 5mC and 5hmC play key roles in the regulation of gene expression. 5mC is associated with transcriptional repression, methylation patterns are established during cell fate determination constraining cell's transcriptional programs. The transition from 5mC to 5hmC occurs in actively transcribed genes and at active or lineage-specific enhancers, serving as both a biomarker and functional DNA modification impacting tissue-specific gene expression.

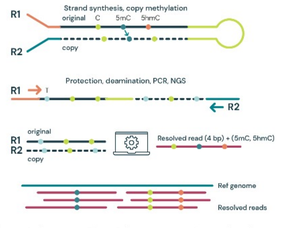


Figure 1. duet evoC is a 6-base sequencing technology that reads all four canonical bases plus 5mC and 5hmC[1].

The value of a 6-base genome has been demonstrated via improved detection of early stage colorectal cancer using 5mC and 5hmC (Puddu et al, 2026). Use of circulating cell-free DNA (cfDNA) for the early detection of tumour-derived fragments has progressed substantially leveraging multimodal data integration, incorporating orthogonal features such as genomic, epigenomic, and fragmentomic biomarkers. Here we show the duet evoC 6-base solution:

- Simultaneously captures genetics, epigenetics and fragmentomics in a single workflow, in a single dataset
- Enables comprehensive characterization of tumour-derived cfDNA using conventional genetic mutational analysis
- Provides new opportunities for biomarker discovery by unlocking the ability to distinguish 5mC from 5hmC
- By delivering complete genetics and epigenetics in a single dataset it allows the discovery of more powerful classifiers for the detection of cancer from healthy cfDNA

Taken together duet evoC 6-base data substantially improves analytical sensitivity for ctDNA detection.

2. Methods

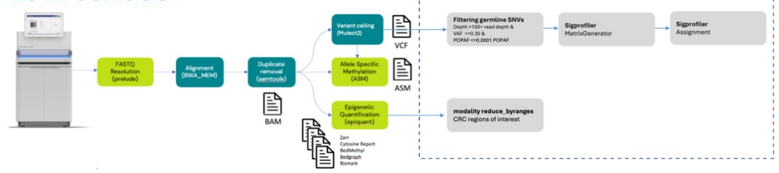


Figure 2. biomodal duet pipeline overview and tertiary analysis steps to obtain features for combined genetic and epigenetic analysis.

As part of the duet evoC solution an HPC and cloud-compatible software pipeline takes standard short-read 4-base FASTQ files and processes them using the duet read resolution logic to produce a 6-base FASTQ output. The duet pipeline also conducts quality filtering and trimming before alignment to the reference genome to produce a 6-base BAM and methylation quantification files.

Mutation calling, mutation selections and mutational signatures generation

Somatic mutation calling was performed using Mutect2 as part of the biomodal duet pipeline. Variant calls were subsequently subjected to stringent filtering to minimise the inclusion of residual germline events. Specifically, single nucleotide variants (SNVs) were retained if they exhibited a variant allele frequency (VAF) <= 0.35, sequencing depth >=15X, and a population allele frequency <= 1×10^{-4} based on gnomAD.

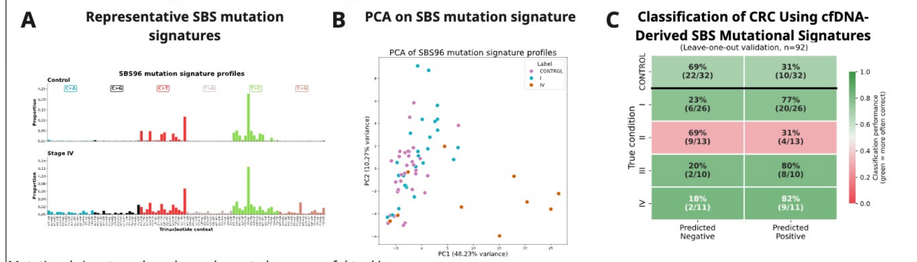
Filtering thresholds were optimised empirically to maximise discrimination between cancer and control samples. This was achieved using an elastic net–regularised logistic regression framework with nested cross-validation, enabling robust feature selection while mitigating overfitting.

Mutational signatures were inferred using SigProfiler. For each sample, single base substitution (SBS) profiles were generated in the 96-channel context (SBS96), capturing the trinucleotide sequence context of each mutation and enabling downstream analysis of mutational processes.

Epigenetic 5mC and 5hmC features

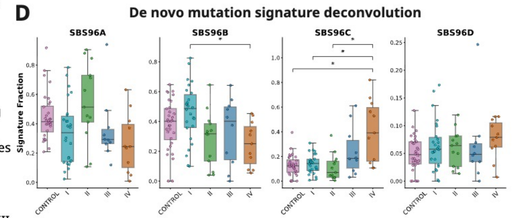
Leveraging the ability of duet evoC to differentiate 5mC and 5hmC from cfDNA, features were calculated for each sample in predefined regions of interest, including loci previously identified as clinical biomarkers and regions exhibiting epigenetic alterations in colorectal cancer (CRC) versus control. Repetitive elements were obtained from RepeatMasker and further filtered to retain highly conserved repeat motifs. Methylation values in the regions of interest were defined as the mean methylation fraction within each region, calculated as the proportion of reads supporting methylation relative to the total read count.

3. The 6-base genome enables mutation signatures analysis that distinguish healthy controls from cancer cfDNA



Mutational signatures have been shown to be a powerful tool in the arsenal for analysis of cfDNA and detection of ctDNA. The complete genetic information provided by the 6-base genome enables the extraction of single base subsitution (SBS) profiles from a healthy and a stage IV CRC cfDNA sample **(A)**. In these samples the importance of duet evoC's accurate C>T calling is highlighted, which is a limitation of other epigenetic sequencing approaches.

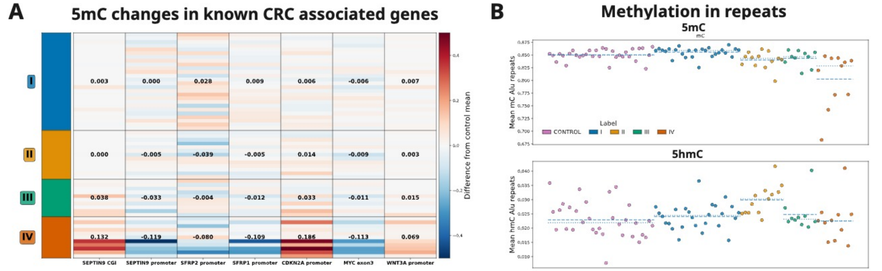
A principal component analysis (PCA) of SBS96 mutational profiles derived from cfDNA **(B)** for controls and CRC stages I and IV suggests genetic biomarkers are most useful at later cancer stages. This observation is confirmed using a classification approach where the highest performance is observed is stages III and IV CRC **(C)** but with a notable failure to classify stage II CRC. The classifier was developed using elastic net–regularised logistic regression model ($\alpha = 0.5$, $\lambda = 0.1$; scikit-learn) and evaluated using leave-one-out cross-validation (LOOCV). Input features comprised SBS96 trinucleotide mutation proportions, which were median-imputed, standardised, and projected onto 15 principal components within each fold to prevent data leakage. Class weights were balanced to account for unequal group sizes. Performance is reported as the proportion of correctly classified samples per group. All CRC stages except stage II were more frequently classified as positive than negative, while controls were predominantly classified as negative.



Mutational signatures can provide some information on the mutational processes that underpin them, by comparison to profiles with known characteristics, and so provide limited biological insight. In this cohort de novo mutational signature extraction and refitting performed using SigProfiler identified four distinct mutational processes within the cohort **(D)**. These data further demonstrate that duet evoC provides complete genetic alongside epigenetic and fragmentomic information from cfDNA

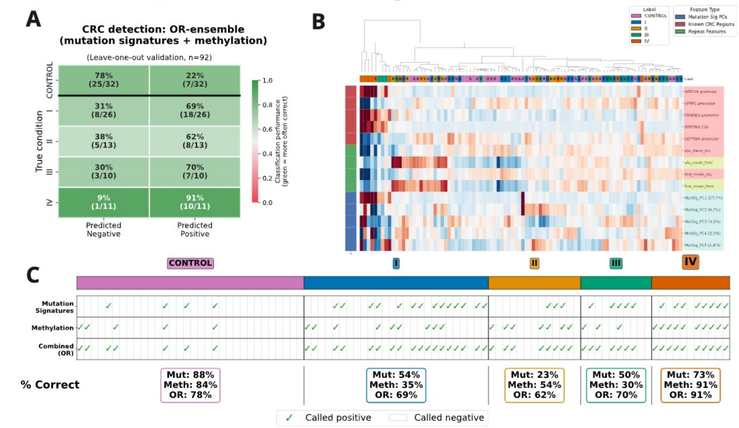
Differences in signature contributions between disease stages were assessed using two-sided Mann–Whitney U tests across all pairwise comparisons, with Benjamini–Hochberg correction for multiple testing (FDR < 0.05). The SBS96C-associated process was significantly enriched in stage IV samples relative to controls and stages I–II, while SBS96A and SBS96B showed a non-significant decrease from controls to stage IV.

4. 5mC and 5hmC levels vary in known cancer marker regions



Given duet evoC provides complete epigenetic information, distinguishing 5mC from 5hmC, we examined genes or regions known to be associated with CRC to understand how their methylation state changed between the stages of cancer. First looking at 5mC we see that there is a general increase in selected genes as cancer progresses **(A)**. Each column represents a genomic region, and each tile (in each box) corresponds to an individual sample, showing the difference in mean methylation relative to controls. Hypomethylation of repeat elements like Alu repeat transposable elements is a hallmark of cancer leading to dysregulation and an increase in their activity. In **(B)** we see the value of distinguishing 5mC from 5hmC, there is a modest decrease in 5mC is observed in stages II–IV compared to controls, while 5hmC shows a distinct trend with an increase in stage II relative to other stages and no loss in stages II-IV. Colours indicate sample groups (control, stages I–IV), dotted lines represent the median value for each group, dashed lines represent mean.

5.Combining 6-base genetic and epigenetic (5mC and 5hmC) information improves ctDNA detection



To better evaluate the utility of combining mutation, 5mC and 5hmC features a classifier was built **(A)** which outperformed the mutation-only classifier, improving performance at all stages, with the improvement at stage II CRC most notable. An ensemble elastic net–regularised logistic regression model was used and evaluated using leave-one-out cross-validation (LOOCV). The same features were used in hierarchical clustering **(B)** using genetic (light teal), 5mC (coral) and 5hmC (green) features. Clustering was conducted using Euclidean distance and Ward linkage. Samples with low tumour fraction (< 0.05, estimated by IchorCNA) are indicated by '×'. Finally the utility of mutation-only, methylation-only or combined classifiers was explored **(C)** through an ensemble model (elastic net-regression models with individual target sensitivity of 85% and an OR statement providing the final call decision). Higher sensitivity is achieved via the combination of the two feature types, with mutation signatures correctly identifying more stage I and stage III CRC samples, while methylation correctly identified more stage II and stage IV samples. Taken together these data demonstrate that combining mutation, 5mC and 5hmC information if more powerful than either alone for the detection of cancer cfDNA.

6. Conclusions

We demonstrate the value of integrating genetic and epigenetic information from a single cfDNA assay using the duet evoC technology, which enables simultaneous extraction of complete genetics and complete epigenetics from the same sample.

- Mutation signatures alone enable robust classification of the majority of advanced-stage samples and allow discrimination of a subset of early-stage cases from healthy controls.
- Incorporating epigenetic features further improves sample stratification, as reflected by clearer separation in clustering analyses and improved classification performance in supervised models.
- Genetic and epigenetic signals provide complementary information, and that their integration enhances the detection of cancer-associated patterns in cfDNA.
- This multimodal approach represents a promising strategy for blood-based applications and supports the potential for more sensitive and robust, non-invasive cancer detection in clinical settings.

In posters #123 and #7844 we present data showing the power of fragment-based classification approaches and combining genetic with 5mC and 5hmC features to enhance liquid biopsy performance. In combination with the data presented here they reinforce the ability of the 6-base genome, provided through duet evoC, to extract the most complete and powerful set of biomarkers that can improve early detection, therapy selection and detection of recurrence of cancer in a liquid biopsy setting.

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. Puddu F. et al. 5-methylcytosine and 5-hydroxymethylcytosine are synergistic biomarkers for early detection of colorectal cancer, *Commun Med* **6**, 15 (2026).
3. Islam S. et al. Uncovering novel mutational signatures by *de novo* extraction with SigProfilerExtractor. *Cell Genomics*. 2, 11 (2022).
4. Wasserkort, R. et al. Aberrant septin 9 DNA methylation in colorectal cancer is restricted to a single CpG island. *BMC Cancer* **13**, 398 (2013).
5. López-Moyado, et al. Paradoxical association of TET loss of function with genome-wide DNA hypomethylation, Proc. Natl. Acad. Sci. 116, 34 (2019).