

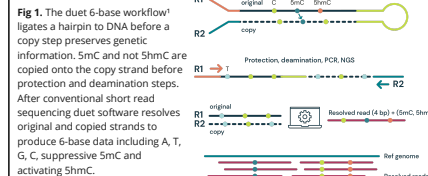
6-base genome sequencing reveals functional genomic readouts in otherwise intractable clinical samples

Craig Fishman¹, Christine Hipsky¹, Aurelie Modat¹, Robert Crawford¹, Luke Sarre¹, Annelie Johansson¹, Angela Simeone¹, Fabio Puddu¹, Tom Charlesworth¹, Robert Osborne¹
¹biomodal Ltd, Cambridge, UK

bioModal

1. Introduction

Functional genomic profiling of clinical specimens is often limited in sample types where RNA or chromatin is degraded, or chemically compromised and chromatin architecture is not preserved, including cell-free DNA (cfDNA) and formalin-fixed paraffin-embedded (FFPE) tissue. DNA is more tractable in these settings but conventional epigenetic sequencing approaches provide incomplete information because they either collapse cytosine modifications into a single modC signal or ignore 5hmC completely, limiting inference of regulatory state information whilst also compromising ability to detect important C-T variants. Measuring genomics, 5mC and 5hmC from the same DNA sample may therefore provide a route to infer functional genomic information where direct RNA or chromatin profiling is not feasible.



2. 6-base sequencing produces high quality data from clinical FF and FFPE samples

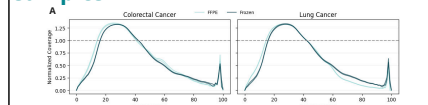


Fig 2. Fresh Frozen (FF) and FFPE clinical CRC and lung cancer samples were processed through the duet to yield high quality and concordant 6-base data.
A. The quality of 6-base data, judged by evaluating GC bias, is robust to formalin damage with equivalent curves for matched FF and FFPE samples, indicating that 6-base data between these sample types is equivalent and can be integrated in downstream analysis.

Methylation correlations	Single CpG 5mC	100kb 5hmC	EM seq v2 modC
CRC FFPE vs FF	0.938	0.636	0.902
Lung Cancer FFPE vs FF	0.890	0.864	0.856

Table 1. Pearson correlations show highly concordant 5mC and 5hmC data across clinical FF and FFPE samples, with higher correlation than EM-seq v2 modC calls, reinforcing the high quality of 6-base data even from FFPE samples.

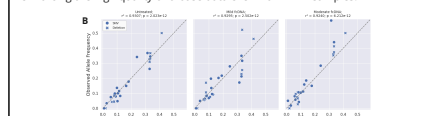


Fig 2B. Observed vs Expected Variant Allele Frequencies (VAFs) for formalin-compromised DNA standards
Expected allele frequencies were obtained from the manufacturer (Horizon Discovery) and compared to observed allele frequencies. A high correlation with $R^2 > 0.92$ up to moderate formalin damage indicates that 6-base data supports accurate SNV calling from FFPE samples.

3. 5hmC reflects gene activity and tissue-specific expression

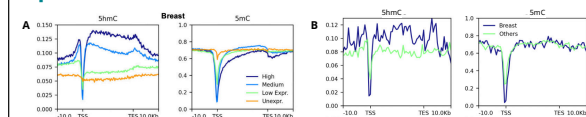


Figure 3A. (A) The relationship between 5hmC and 5mC and gene expression in breast tissue using matched 6-base and RNA-seq data from the same sample. High, medium, low, and unexpressed genes using thresholds were grouped and the 5mC and 5hmC level plotted across scaled regions centred on the gene body. 5hmC is most abundant in the gene bodies of highly expressed genes and high gene expression is associated with low promoter 5mC and 5hmC. **(B)** Comparison of 5mC and 5hmC levels specifically expressed in breast tissue compared to those specifically expressed in other tissues.

Unlike conventional methylation approaches that measure a conflated modC signal, the duet technology independently resolves 5mC and 5hmC, preserving the distinct biological information carried by each modification. Highly expressed genes exhibit elevated gene-body 5hmC and reduced 5mC (**Figure 3A**). The separation between different levels of expression is far more distinct for 5hmC than for 5mC, even with a lower dynamic range. Together these observations suggest that the ability to distinguish 5mC from 5hmC could provide significant value in predicting functional genomic readouts like gene expression over traditional modC methylation sequencing data. These opposing patterns would be obscured if the two modifications were combined as modC.

Furthermore, tissue-specific genes more clearly show elevated gene-body 5hmC relative to genes expressed in other tissues than 5mC (**Figure 3B**), supporting 5hmC as a powerful biomarker of tissue-specific gene activity. These observations are consistent with recent work that demonstrated combined 5mC and 5hmC measurements can be used to infer gene expression in tissue from cfDNA (1).

4. Complementary 5mC and 5hmC signals support gene-expression inference

Given 5hmC positively correlates with gene expression, whereas promoter 5mC negatively correlates with gene expression, the combination of these two features could be a simple and powerful proxy for gene expression in samples, like cfDNA and FFPE, where RNA-seq is intractable. In cfDNA, combining promoter 5mC and gene body 5hmC as a composite methylation z-score has been shown to accurately infer gene expression changes in the tumour in response to treatment². We used matched RNA-seq and 6-base data from breast, colon, leukocyte, liver and lung tissues to evaluate this approach in tissue samples.

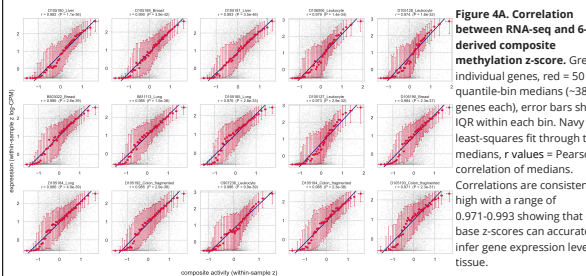


Figure 4A. Correlation between RNA-seq and 6-base derived composite methylation z-score. Grey = individual genes, red = 50 quantile-bin medians (~380 genes each), error bars show IQR within each bin. Navy line = least-squares fit through the 50 medians, r values = Pearson correlation of medians. Correlations are consistently high with a range of 0.971-0.993 showing that 6-base z-scores can accurately infer gene expression levels in tissue.

Tissue	Top ARH54 tissue signature
Liver	LIVER, HEPATOCYTE
Leukocyte	PERIPHERAL BLOOD, NEUTROPHIL, GRANULOCYTE
Colon	COLON, ILEUM, SMALL INTESTINE, COLONIC MUCOSA
Lung	LUNG, ALVEOLAR CELL TYPE II (+ macrophage)
Breast	ADIPOSE / SUBCUTANEOUS ADIPOSE / OMENTUM, BREAST

Table 1. The top 150 expressed genes, ranked using the 6-base z-score, were taken to assess the tissue-specificity of the methylation-derived gene expression readout. Genes were compared against tissue signatures based on RNA-seq data³ which validated that predicted highly expressed genes reflect the known biology of each tissue.

5. Inference of regulatory element status and chromatin accessibility with 6-base data

Chromatin accessibility and enhancer status provide insights into genome organisation and regulation that can be used to understand the impact of genetic changes, especially in non-coding regions of the genome. However either approach, using either ATAC-seq or histone-specific pulldown approaches, require intact chromatin which can be challenging in FFPE samples or absent in cfDNA. Therefore alternative approaches are required to obtain functional readouts in these challenging sample types.

Figure 5A. Classification of enhancer state with 6-base data

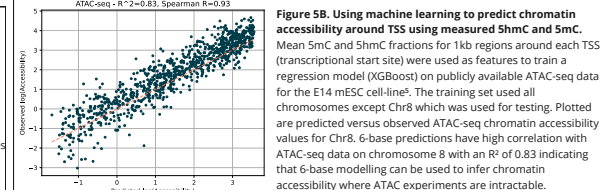
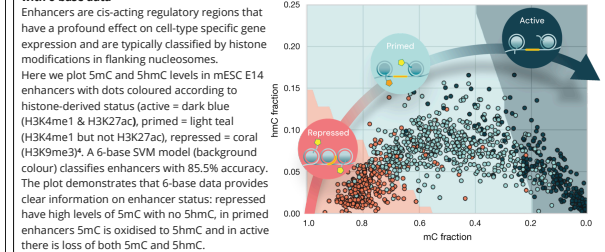


Figure 5B. Using machine learning to predict chromatin accessibility around TSS using measured 5hmC and 5mC. Mean 5mC and 5hmC fractions for 1kb regions around each TSS (transcription start site) were used as features to train a regression model (XGBoost) on publicly available ATAC-seq data for the E14 mESC cell-line⁴. The training set used all chromosomes except Chr8 which was used for testing. Plotted are predicted versus observed ATAC-seq chromatin accessibility values for Chr8, 6-base predictions have high correlation with ATAC-seq data on chromosome 8 with an R^2 of 0.83 indicating that 6-base modelling can be used to infer chromatin accessibility where ATAC experiments are intractable.

6. Conclusions

The investigation of genetic changes, especially in non-coding regions, often requires a host of methods that demand intact chromatin and high quality RNA which are intractable in precious clinical samples such as FFPE and cfDNA. With the advent of 6-base genome, which provides not only complete genetic information but also suppressive 5mC and activating 5hmC, it is now possible to access functional readouts even where RNA or intact chromatin are unavailable:

- The duet workflow provides reproducibly high genetic and epigenetic quality data from clinical FFPE and cfDNA samples
 - Gene body 5hmC is a clear and stable marker of gene expression
 - Combining information on promoter 5mC and gene body 5hmC enables accurate inference of gene expression programmes
 - 5mC and 5hmC levels can be used to accurately infer enhancer status
 - Similarly, simple machine learning models of 5mC and 5hmC levels around the TSS infers chromatin accessibility
- The 6-base genome thus offers a platform not only for combined genetic and epigenetic biomarker discovery but also a singular dataset that extends from discovery to functional interpretation.

7. References

1. Fullgrabe J, et al. Simultaneous sequencing of genetic and epigenetic bases in DNA. Nat Biotechnol. 41(10):1457-1464 (2023).
2. Savelkoul AG, et al. Integrative dual-cDNA 5mC/5hmC methylomics and clonal reconstruction infer tumor transcription and resistance phenotypes in metastatic prostate cancer. Cell Reports Medicine. 3(25).
3. <https://arxiv.org/abs/2008.00000>
4. Cruz-Molina S, Resnault P, Tebbat C, et al. PRIC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. Cell Stem Cell. 27(1):203-218 (2021).
5. Moresco M, van den Brand T, L.H. Teunissen H, Davies J, de Wit E. Pioneer activity distinguishes activating from non-activating SOX2 binding sites. EMBO J. 33(24):2021-2031 (2014). doi:10.1525/embo.2013.21315

