Inferring gene regulation from 6-base sequencing data in multiple tissues

Mark Consugar[1], Annelie Johansson[1], Ermira Lleshi[1], Mark S. Hill[2], Jack Monahan[3], Fabio Puddu[1], William Stark[4], Jean Teyssandier[1], Aurélie Modat[1], Robert Crawford[1], Tom Charlesworth[1], Robert J Osborne[1], Páidí Creed[1]

1. biomodal Ltd, The Trinity Building, Chesterford Research Park, Cambridge, UK
2. Isomorphic Labs, London, UK
3. Hurdle, UK
4. Ride Therapeutics, Cambridge, UK

biomodal

1. Introduction

5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), also known as the fifth and sixth DNA bases of the genome, are known to have different distributions in different tissues and can act as tissue-specific fingerprints.

Furthermore, 5mC and 5hmC play key roles in the regulation of gene expression. Specifically, 5mC is associated with transcriptional repression, with patterns of methylation established during cell fate determination constraining the transcriptional programs in cells. 5mC is oxidised by the TET enzymes to 5hmC in the gene bodies of actively transcribed genes and at active or lineage-specific enhancers, serving as both a biomarker and functional DNA modification impacting tissue-specific gene expression.

Here we apply 6-base sequencing using duet evoC, a multiomic solution capable of simultaneously detecting all four canonical DNA bases along with 5mC and 5hmC, to comprehensively profile the 6-base genome of multiple tissues. We use this dataset to demonstrate that even where there are low levels of 5hmC compared to 5mC, there is significant value in distinguishing the two methylation biomarkers with 5hmC a powerful marker for tissue-specific genes that drives their expression.

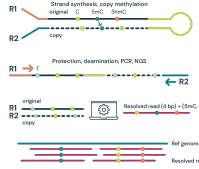


Figure 1. duet evoC is a 6-base sequencing technology that reads all four canonical bases plus 5mC and 5hmC[1].

2. Methods

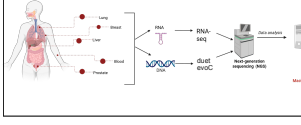


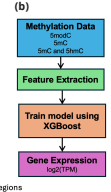
Figure 2. Whole-genome 6-base data was generated using duet evoC alongside matched RNA-seq data from fresh frozen tissues representing five distinct human organs, collected from 11 healthy donors. In total, 32 samples were processed, comprising 16 matched DNA and RNA extracts from 4 leukocyte, 3 breast, 3 prostate, 3 lung, and 3 liver specimens. Sequencing used the NovaSeq 6000 platform, achieving ~25x coverage for duet evoC libraries and >100X depth for RNA-seq libraries.

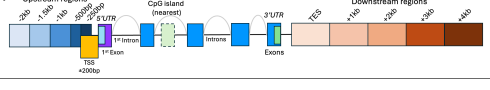
5. Machine learning model

Figure 5. Genomic features included in the machine learning model

(a) To predict gene expression from 6-base sequencing data, mean 5mC and 5hmC fractions (methylation features) were summarised over genomic regions located around a gene, spanning upstream regions including promoter (-2kb to + 1kb relative to transcription start site, TSS), gene bodies (exons and introns), 3'UTRs, and downstream regions (+1kb to +4kb from transcription end site, TES). CpG island methylation was assessed for the nearest regulatory island.

(b) Our machine learning model integrates methylation data to generate ~50-60 features per gene and trains the model using XGBoost to predict gene expression log2(TPM) values.





3. 6-base data correlates with gene expression

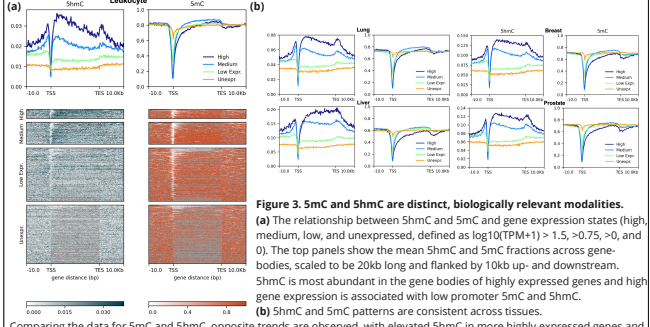


Figure 3. 5mC and 5hmC are distinct, biologically relevant modalities.

(a) The relationship between 5hmC and 5mC and gene expression states (high, medium, low, and unexpressed, defined as log10(TPM+1) > 1.5, >0.75, >0, and 0). The top panels show the mean 5hmC and 5mC fractions across gene-bodies, scaled to be 20kb long and flanked by 10kb up- and downstream. 5hmC is most abundant in the gene bodies of highly expressed genes and high gene expression is associated with low promoter 5mC and 5hmC.

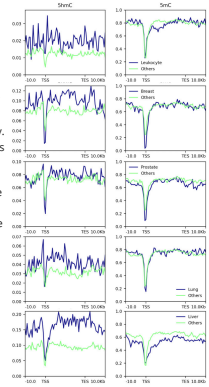
(b) 5hmC and 5mC patterns are consistent across tissues.

Comparing the data for 5mC and 5hmC, opposite trends are observed, with elevated 5hmC in more highly expressed genes and a concomitant decrease in 5mC. Were 5mC and 5hmC to be conflated into a traditional 5modC readout these opposing trends would be obscured and would cancel one another, reducing the biological signal. Furthermore, 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 5modC methylation sequencing data.

4. Tissue-specific gene expression is driven by 5hmC

Figure 4. 5mC and 5hmC in tissue-specific genes

Tissue-specific gene expression was defined as high or medium expression for a specific tissue, and low or unexpressed for all other tissues. This identified 96, 52, 99, 69, and 117 uniquely expressed genes for Leukocyte, Breast, Prostate, Lung, and Liver, respectively. Most consistent is the increase in 5hmC levels in the gene body of tissue-specific genes relative to the set of genes which are specific to the other tissues, with an average increase of +41%. The relative increase in 5hmC is largest in Liver (+78%). Around the TSS, there is a decrease in 5mC in tissue-specific genes, with an average of -16%. The largest relative 5mC decrease is seen in Breast (-35%). These data reinforce the value of differentiating 5mC from 5hmC. 5mC information, which is the dominant component of 5modC data, appears to be less able to distinguish tissue-specific genes, whereas 5hmC provides a much clearer signal.



6. 6-base data predicts gene expression

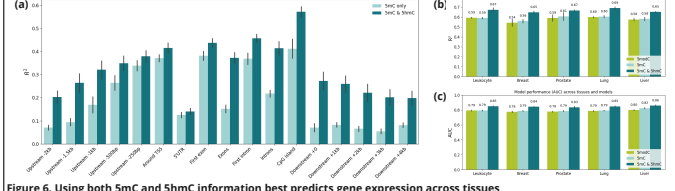


Figure 6. Using both 5mC and 5hmC information best predicts gene expression across tissues

(a) Assessing the contribution of single features to gene expression prediction using correlation analysis reveals a consistent increase in performance when using both 5mC and 5hmC information compared to 5mC alone. We then trained a model on all methylation data features to predict gene expression log2(TPM) values. **(b)** Robust predictive performance was observed with significantly increased performance when using both 5mC and 5hmC information, compared to 5modC or 5mC only. R^2 values ranged from 0.51-0.61 for 5modC, 0.54-0.63 for 5mC only, and 0.64-0.71 for models including both 5mC and 5hmC. We saw a relative increase in R^2 of between 10% and 16%, with the largest gain in Breast and the smallest in Prostate. **(c)** Similar gains in AUC were observed in models trained to predict 4 categorical gene expression states.

These data robustly and comprehensively demonstrate the additional value derived from 6-base data when compared to either a 5modC or 5mC only readout. 6-base data consistently provides more accurate information on the functional genomics of a tissue, enabling the inference and prediction of gene expression.

7. Conclusion

We have shown that patterns of 5hmC and 5hmC provide insight into tissue-specific gene expression with the ability to better distinguish high and low expressed genes than 5mC or 5modC data. Although most tissues have far lower levels of 5hmC than 5mC, 5hmC appears to be a clearer biomarker of highly expressed, tissue specific genes and so plays a critical functional role in the genome.

Building on this we used machine learning approaches to robustly demonstrate the generalisability of this observation across tissues. Models using features derived from 5mC and 5hmC are consistently more predictive of gene expression than using 5mC alone or 5modC.

Together these data illustrate the potential for using 6-base sequencing, which provides accurate measurements of both 5mC and 5hmC, to characterize, understand and predict functional genomics such as transcriptomic programs of cells directly from DNA. Notably, this could enable transcriptional profiling in cases where RNA is not accessible or difficult to extract, for example in cell-free DNA or FFPE-preserved tissue samples. This powerful capability unlocks the ability to derive mechanistic insight from a 6-base readout across diverse sample types.

8. References

1. Füllgrabe J. et al. Simultaneous sequencing of genetic and epigenetic bases in DNA. Nat Biotechnol. 2023 Oct;41(10):1457-1464.

biomodal