Inferring gene regulation from 6-base sequencing data

William Stark[1], Mark Hill[1], Jack Monahan[1], Jean Teyssandier[1], Chenfu Shi[1], **Nicola Wong**[1], Hugo Sepulvega[3], Isaac Lopez-Moyado[2], Nicholas Harding[1], Anjana Rao[2], Páidí Creed[1]

1 biomodal Ltd, The Trinity Building, Chesterford Research Park, Cambridge, UK
2 Division of Signaling and Gene Expression, La Jolla Institute for Immunology, La Jolla, CA, USA
3 Universidad Andrés Bello, Santiago, Chile

biomodal

1. Introduction

DNA methylation, an epigenetic modification, plays a key role in the regulation of gene expression. Specifically, the addition of a methyl group to the 5th carbon of cytosine (5mC) is broadly associated with transcriptional repression, with patterns of methylation established during cell fate determination constraining the transcriptional programs in cells. Demethylation via oxidation of 5-methylcytosine to 5-hydroxymethylcytosine (5hmC) is performed by TET enzymes and is reflected by higher levels of 5hmC in the gene bodies of actively transcribed genes and at active or lineage-specific enhancers.

Disentangling the roles of these two distinct modifications in gene regulation has been constrained by technological limitations, with most sequencing approaches conflating the two modifications into a single measure representing 5mC or 5hmC. Recent developments in sequencing technologies have enabled base-resolved simultaneous measurement of 5mC and 5hmC. Utilising these developments, we have generated data-sets with whole-genome 5mC and 5hmC measurements paired with RNA-sequencing data, across several different cell and tissue types. We show that we can see distinct patterns of 5mC and 5hmC across different cell types, that patterns of 5mC and 5hmC reflect tissue-specific gene expression, and that machine learning models can be trained to predict gene expression from 5mC and 5hmC.

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

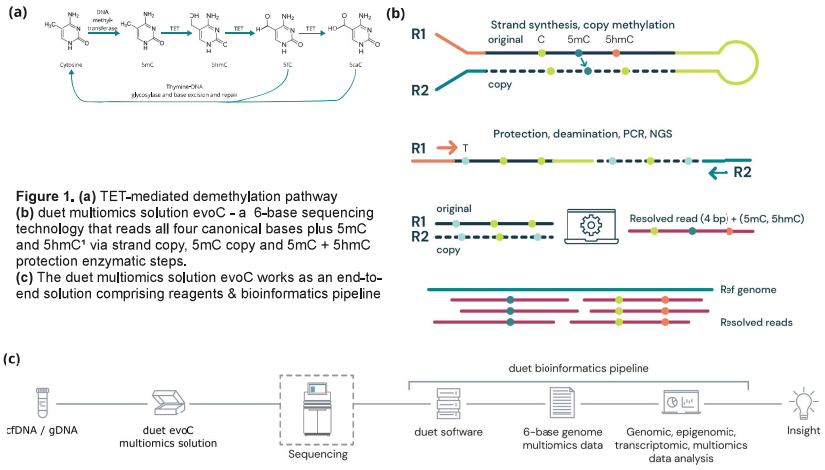


Figure 1. (a) TET-mediated demethylation pathway
(b) duet multiomics solution evoC - a 6-base sequencing technology that reads all four canonical bases plus 5mC and 5hmC[1] via strand copy, 5mC copy and 5mC + 5hmC protection enzymatic steps.
(c) The duet multiomics solution evoC works as an end-to-end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

3. Tissue-specific 5mC and 5hmC and gene expression

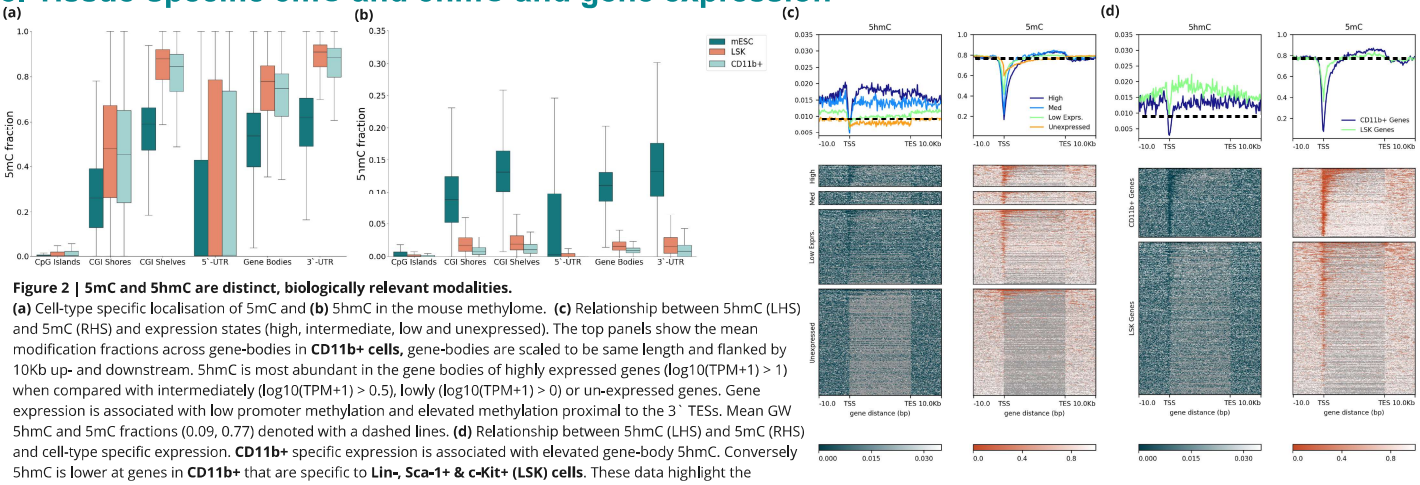
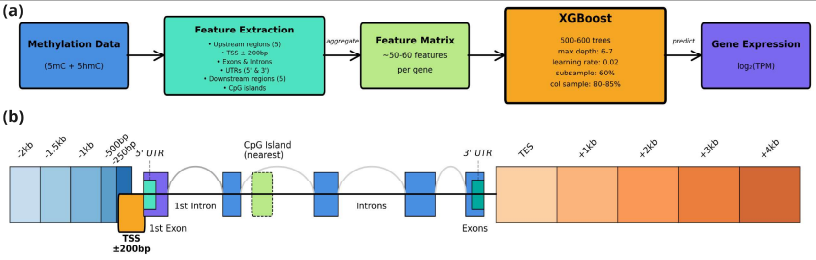


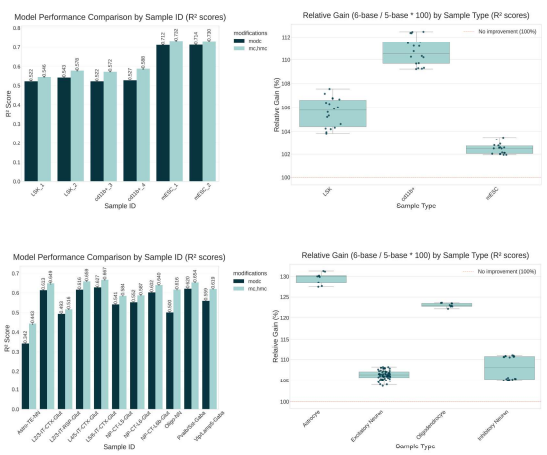
Figure 2 | 5mC and 5hmC are distinct, biologically relevant modalities.
(a) Cell-type specific localisation of 5mC and **(b)** 5hmC in the mouse methylome. **(c)** Relationship between 5hmC (LHS) and 5mC (RHS) and expression states (high, intermediate, low and unexpressed). The top panels show the mean modification fractions across gene-bodies in **CD11b+ cells,** gene-bodies are scaled to be same length and flanked by 10Kb up- and downstream. 5hmC is most abundant in the gene bodies of highly expressed genes (log10(TPM+1) > 1) when compared with intermediately (log10(TPM+1) > 0.5), lowly (log10(TPM+1) > 0) or un-expressed genes. Gene expression is associated with low promoter methylation and elevated methylation proximal to the 3` TESs. Mean GW 5hmC and 5mC fractions (0.09, 0.77) denoted with a dashed lines. **(d)** Relationship between 5hmC (LHS) and 5mC (RHS) and cell-type specific expression. **CD11b+** specific expression is associated with elevated gene-body 5hmC. Conversely 5hmC is lower at genes in **CD11b+** that are specific to **Lin-, Sca-1+ & c-Kit+ (LSK) cells**. These data highlight the distinctive dynamics of 5mC and 5hmC at TSSs, gene bodies and TESs for genes with different expression states.

4. Machine Learning model architecture

(a), Our pipeline integrates 5mC and 5hmC methylation data across diverse genomic loci to generate ~50-60 features per gene. An ensemble XGBoost classifier with 500-600 trees predicts transcript abundance (log_2TPM) using optimized hyperparameters (max depth: 6-7, learning rate: 0.02, subsampling: 60-85%).
(b), Methylation features are extracted from spatially-defined genomic intervals spanning promoter regions (-2kb to +1kb relative to TSS), gene bodies (exons and introns), 3' UTRs, and downstream sequences (+1kb to +4kb from transcription end sites). CpG island methylation status is assessed for the nearest regulatory island within each gene's genomic vicinity. This multi-scale approach captures both proximal regulatory elements and distal chromatin modifications that influence transcriptional output.



5. Predict RNA-seq from 6-base data



Cross-sample validation demonstrates robust predictive performance with R^2 values ranging from 0.52-0.73 across diverse cell types and tissues. The 5mC+5hmC methylation feature set consistently outperforms 5mC-only models (modC), with mean improvements of 4-12% across stem cell lines (LSK, mESC), differentiated cells (cd11b+), and primary tissues. Box plots stratified by sample type reveal that astrocytes and oligodendrocytes exhibit the highest relative performance gains (25-30%), while excitatory neurons show more modest improvements (5-7%). The consistent performance enhancement across cell types validates the biological relevance of hydroxymethylation in transcriptional regulation prediction.

6. Conclusion

We have shown that patterns of 5hmC and 5hmC provide a view on tissue-specific gene expression. In particular, in Cd11b+ myeloid cells from mouse, levels of 5hmC in the gene body were shown to track with the level of gene expression and to distinguish genes which were uniquely expressed in Cd11b+ cells relative to LSK cells (a type of hematopoietic stem cell from which Cd11b+ cells differentiate).
Building on this we showed that we can build models to predict gene expression using features derived from 5mC and 5hmC, with models based on this feature set consistently outperforming models with features derived from modified Cytosine data (which does not discriminate between 5mC and 5hmC). This illustrates the potential for using 6-base sequencing, which provides accurate measurements of both 5mC and 5hmC, to characterize the 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

7. References

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

biomodal