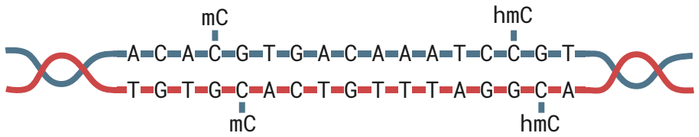
Accurate and simultaneous sequencing of genetics and epigenetics in DNA

Páidí Creed, David Currie, Nicholas Harding, Casper K Lumby, Jack Monahan, David M Morley, Fabio Puddu, Jamie Scotcher, Rosie Telford, Jean Teyssandier, Michael Wilson, Shirong Yu, Joanna D Holbrook

Cambridge Epigenetix Ltd, UK

1. Introduction

There is more to DNA than A, C, G and T. Epigenetics plays a causal role in cell fate, ageing and disease development. Methylated cytosines, such as 5mC and 5hmC, represent important biomarkers and are informally considered the 5th and 6th letters in the genetic alphabet.



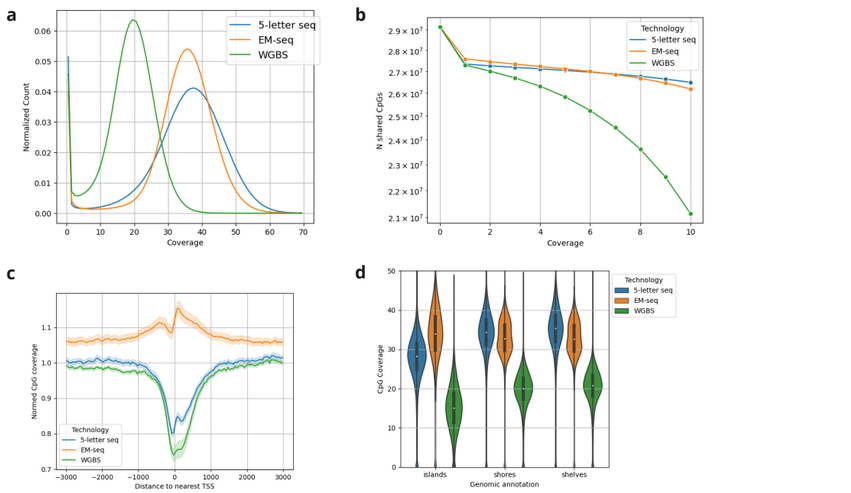
Constrained to measuring four states of information, existing NGS-based technologies sacrifice genetic insight (ability to differentiate C and T) for methylation calling.

We have developed a sequencing technology, **5-Letter seq**, that overcomes these challenges and facilitates unique insights arising from reading genetic and epigenetic letters on the same read and at high accuracy.

State	Standard sequencing protocol	Protocol with C→T deamination
1	A	A
2	C/mC/hmC	mC/hmC
3	G	G
4	T	C/T

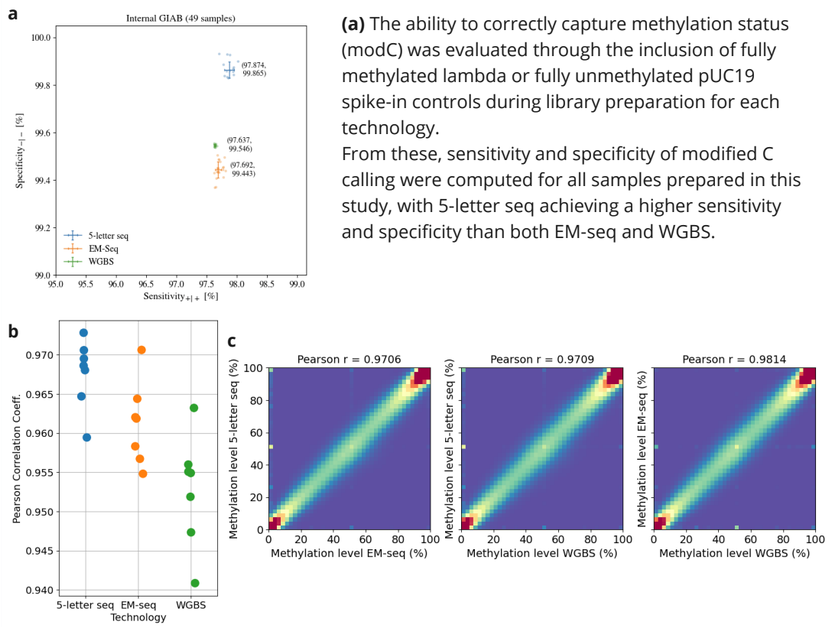
We have created a benchmark dataset, using all 7 Genome in a bottle[1] samples (2x technical replicates per sample). This dataset enables a comparison of the 5-letter seq protocol with two standard methylation sequencing protocols, the xGen™ Methyl-Seq whole genome bisulfite library prep kit (WGBS) and NEBNext® Enzymatic Methyl-seq Kit (EM-seq). See **Poster 520** for a detailed evaluation of SNP calling using 5-letter seq, also using this data-set

2. Uniform and consistent coverage of CpGs



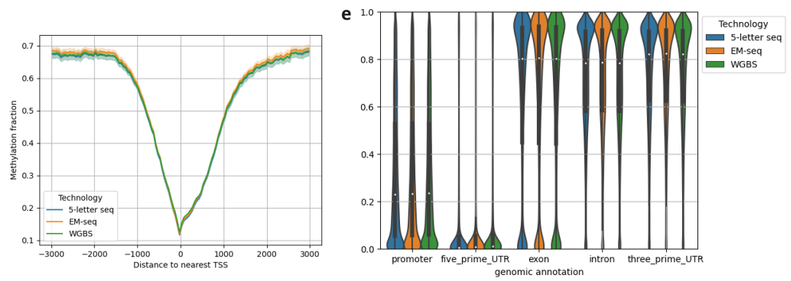
All samples were down-sampled to 36.6X (minimum observed) genomic coverage. Top and bottom strand CpGs were merged, yielding a maximum of ~29 million possible CpG sites. **(b)-(d)** aggregate data from all 14 samples. **(a)** Empirical distribution of CpG coverage for one sample (HG001). **(b)** Empirical cumulative distribution for the number of CpGs achieving a minimum coverage level across all samples with each technology. 5-letter seq and EM-seq cover the same CpGs more consistently than WGBS. **(c)** Normalised mean coverage near Transcription Start Sites. **(d)** Empirical distribution of coverage in CpG shelves, shores, and islands.

3. Accurate modC calls, concordant with WGBS



(a) The ability to correctly capture methylation status (modC) was evaluated through the inclusion of fully methylated lambda or fully unmethylated pUC19 spike-in controls during library preparation for each technology.
From these, sensitivity and specificity of modified C calling were computed for all samples prepared in this study, with 5-letter seq achieving a higher sensitivity and specificity than both EM-seq and WGBS.

Methylation fractions estimated with 5-letter seq achieve good correlation between technical replicates and are concordant with both WGBS and EM-seq. **(b)** Pearson correlation coefficient between all pairs of technical replicates, yielding 7 data points for each technology. All technologies achieved good correlation between technical replicates, with 5-letter seq and EM-seq displayed superior consistency to WGBS. **(c)** Representative heatmaps of correlation for EM-seq, WGBS and 5-letter seq across a replicate of HG001.



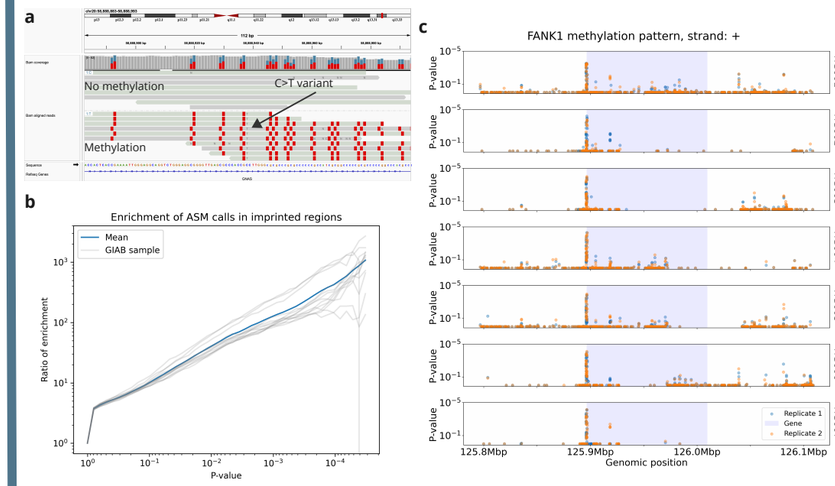
We further validated epigenetic calls by comparing methylation fraction given different genomic annotations. All data presented here is averaged over all samples. **(d)** As expected, we observe decreased methylation near TSS (transcription start sites) in all technologies. **(e)** Violin plots showing the distribution of methylation fraction at different genomic annotations. In all technologies we observe signature low methylation at promoter regions and 5' UTRs.

Data Processing

Adapter sequences were trimmed using CUTADAPT. As advised for protocols using XGen Adaptase technology, WGBS reads were further trimmed for an additional 10 bp on the 3' end of R1 and 10 bp on the 5' end of R2 to remove adaptase sequence introduced during library preparation. No additional trimming was applied to EM-seq or 5-letter seq.
EM-seq and WGBS reads were aligned to the genome using bwa-meth (3-letter alignment) and modified Cytosines were quantified using Methyldackel. 5-letter seq reads were preprocessed using software developed at Cambridge Epigenetix as described in Füllgrabe and Gosal (2023)[2] and subsequently aligned to the genome using BWA (4-letter alignment). Modified Cytosines were quantified using custom software developed at Cambridge Epigenetix.

4. Allele-specific Methylation can be identified in a single sequencing experiment

Allele-specific methylation (ASM) is an important biological phenomenon where differential methylation patterns are observed across the two alleles at a heterozygous variant site. It can result in differential expression of variant alleles and can be associated with tissue-specific activation of oncogenic mutations.
Due to its joint and phased characterisation of genetics and epigenetics, 5-Letter seq is uniquely suited for measuring ASM. In the GIAB samples analysed in this study, we observe ASM at numerous loci across the genome, consistent across replicates and samples. The methodology for calling ASM can be extended to identify Methylation QTLs by examining the association between SNP alleles and nearby methylation structure across a cohort (data not shown).



(a) Schematic of allele-specific methylation. The C>T variant is associated with differential methylation (methylated Cs marked in red, note the offset of the -ve strand). **(b)** Imprinted regions (as identified by independent sources[3]) are highly enriched for ASM calls. **(c)** ASM calls around the TSS of *FANK1*, show significant levels of ASM, consistently across samples and replicates.

References and Communications

1. *Extensive sequencing of seven human genomes to characterize benchmark reference materials.* Zook et al. Scientific Data (2016). https://doi.org/10.1038/sdata.2016.25
2. *Accurate simultaneous sequencing of genetic and epigenetic bases in DNA*, Füllgrabe and Gosal et al., Nature Biotechnology (2023). Available at: https://www.nature.com/articles/s41587-022-01652-0 **(5-letter seq technology paper)**
3. *Genomic map of candidate human imprint control regions: the imprintome.* Jima et al. Epigenetics. (2022) https://doi.org/10.1080/15592294.2022.2091815

Additional CEGX posters at AGBT 2023

- **Poster P414**: *'Profiling genetic and epigenetic changes, at read-level, after cellular rejuve-nation'*, Holbrook et al.
- **Poster P520**: *"Joint genetic and epigenetic sequencing technology leads to improved genetics compared to existing methylation calling methods"*, Lumby et al.

Twitter: @CEGX_news

https://cambridge-epigenetix.com/

cambridge epigenetix