Accurate simultaneous sequencing of genetic and epigenetic bases in DNA

Shankar Balasubramanian[1], Dan Brudzewsky[2], Philippa Burns[2], Tom Charlesworth[2], Páidí Creed[2], Jens Füllgrabe[2], Walraj Gosal[2], Jane D Hayward[2], Joanna D Holbrook[2], Casper K Lumby[2], David J Morley[2], Shirong Yu[2]

[1] University of Cambridge, Cambridge, UK, [2]Cambridge Epigenetix Ltd, authors listed alphabetically by surname, presenting author underlined

INTRODUCTION

DNA contains more information than just the genetic bases G-A-T-C. Epigenetics plays a causal role in cell fate, ageing, response to environment and is disrupted in the very early stages of disease.

The most frequent epigenetic modifications to bases in the human genome are 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), which have distinct functions. Current methods use next-generation sequencing (NGS) to elucidate epigenetic letters after differentially converting bases, dependent on their epigenetic modifications. However, discriminating epigenetic letters is accomplished by sacrificing the ability to distinguish all four genetic letters.

We describe CEGX sequencing technology that has addressed these problems by expanding the number of information states in NGS from four to sixteen which allows direct, digital and phased discrimination of genetic and epigenetic letters on the same read.

CEGX SEQUENCING TECHNOLOGY

CEGX sequencing technology is an end-to-end solution which accesses more information from DNA than previously possible. It enables single base resolution of complete genetic and epigenetic letters from genomic DNA or cell-free DNA (cfDNA) using standard NGS platforms. The associated fully automated and containerised pipeline provides post-sequencing informatics, which produces genetic and epigenetic information at read level in standard file formats compatible with downstream applications.

1A Single Workflow for Genetic and Epigenetic Information

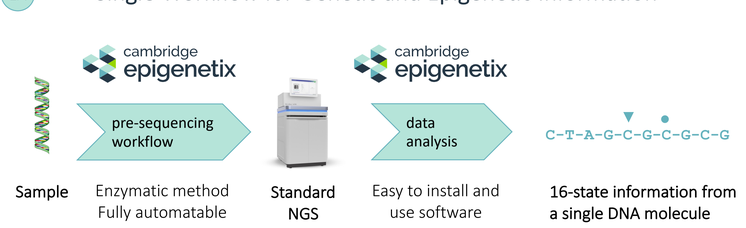
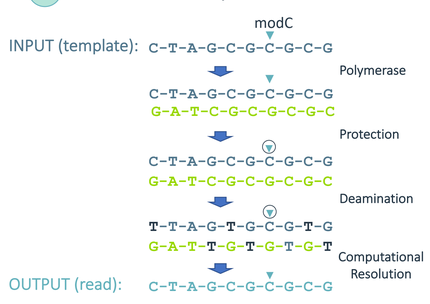


FIGURE 1: CEGX SEQUENCING TECHNOLOGY

A) **The CEGX technology comprises of a single pre-sequencer workflow and post-sequencer software.** It is sequencer agnostic, the first products are optimised for ILMN sequencers with no customisation in combination with a software pipeline that can be deployed on premises on all major cloud providers.

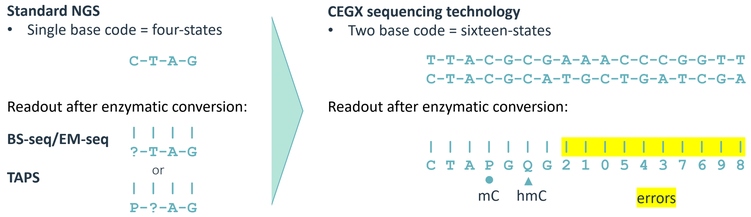
B) **The CEGX sequencing technology generates phased discrimination of genetic and epigenetic letters on the same read**. The gDNA or cfDNA sample undergoes two enzymatic conversion steps for site-specific conversion of unmodified cytosine bases to produce a library for paired-end sequencing. The number of information states is expanded from four-states to sixteen for direct, digital and phased discrimination of genetic and epigenetic letters on the same read and suppression of PCR and sequencing errors.

1B 5-Letter seq workflow



C) Unlike current base conversion methylation detection methods, CEGX sequencing does not sacrifice genetic information. CEGX sequencing digitally discriminates all four genetic letters whilst also capturing epigenetic information. By contrast, standard four-state methods such as bisulphite sequencing (BS-seq), NEBNext Enzymatic Methyl-seq (EM-seq™) or TET assisted pyridine borane sequencing (TAPS) that sacrifice C/T genetic information to capture modified cytosine. Sixteen state encoding also enables suppression of artefacts introduced during sample preparation and sequencing.

1C Genetic and epigenetic letters are digitally discriminated



5-LETTER SEQ DELIVERS BEST-IN-CLASS ACCURACY FOR GENETICS AND EPIGENETICS

The whole genome genetic accuracy at read level and methylome analysis from CEGX 5-Letter seq was compared to BS-seq, EM-seq and standard genome sequencing (ILMN), using 'genome-in-a-bottle' sample NA12878 and ground-truth spike-ins.

The CEGX 5-Letter seq assay delivers more accurate genetic information than BS-seq and EM-seq as it preserves the information about C/T mutations (the most common mutation type). 5-Letter seq results in accurate Phred scores of which 47% are above Q37, this being the highest attainable Phred score for ILMN.

The comparisons also demonstrate that 5-Letter seq has combined higher sensitivity and specificity for detection of modified cytosine (modC) compared to BS-seq and EM-seq. Methylation quantification is highly correlated between BS-seq and 5-Letter seq.

2 Read Level Genetic Accuracy

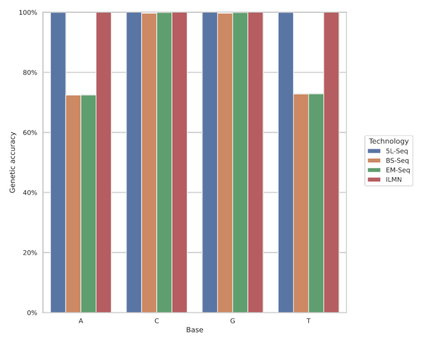
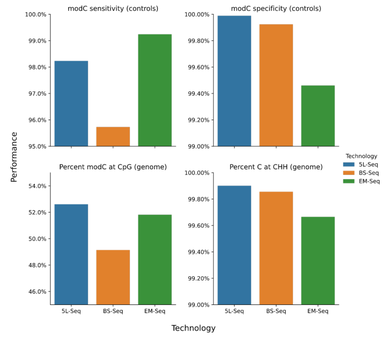


FIGURE 2: GENOME-WIDE READ-LEVEL GENETIC ACCURACY

Genetic accuracy stratified by base type for CEGX 5-Letter seq, BS-seq, EM-seq and ILMN. Input data were restricted to bases with a reported Q-score >= 25. Genomic data from NovaSeq sequencing of NA12878 sample using 80-100ng input DNA.

To account for the underlying sixteen-state process, 5-Letter seq base qualities calibrated according to sequencing platform.

3A Methylation Detection



3B Correlation with BS-seq

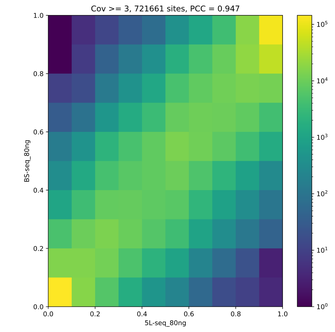


FIGURE 3: ACCURACY OF MODIFIED CYTOSINE DETECTION

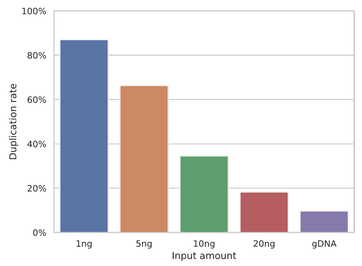
A) **5-Letter seq (blue) combined sensitivity and specificity measured on ground-truth controls is higher than either BS-seq (orange) or EM-seq (green). Consistent with this, 5-Letter seq detects more modified cytosine at CpG sites and less at CHH sites in the genome than either BS-seq or EM-seq.**
Top left panel: sensitivity (modC/modC+C) calls at read level for every CpG in a fully methylated lambda.
Top right panel: specificity (C/modC+C) calls at read level for every CpG is a fully unmethylated pUC19).
Bottom left panel: average methylation levels (modC/modC+C) across all CpGs in the NA12878 genome.
Bottom right panel: average unmethylated levels (C/modC+C) across all CHHs in the NA12878 genome.

B) **Methylation quantification is highly correlated between BS-seq and CEGX 5-Letter seq.**
Correlation heatmap of methylation values at individual CpGs covered by at least 3 reads by both techniques in chromosome 20 of NA12878. 80ng genomic DNA was quantified by CEGX 5-Letter seq (x-axis) or BS-Seq (y-axis).

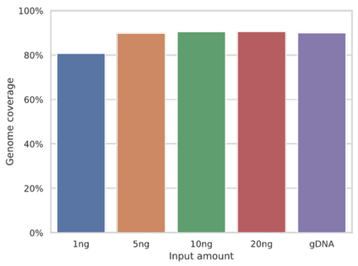
CEGX SEQUENCING TECHNOLOGY IS COMPATIBLE WITH LIQUID BIOPSY

Combined genetic and epigenetic data in liquid biopsies can improve sensitivity of early disease detection and identification of tumour subtype. 5-Letter seq was performed on cell free DNA (cfDNA) from a donor with stage 3 CRC ranging from 1ng-20ng input. Data shows compatibility with the cell free DNA quantity typically available from a standard blood draw, and an example of C>T mutation in the APC gene proximal to a hypermethylated CpG.

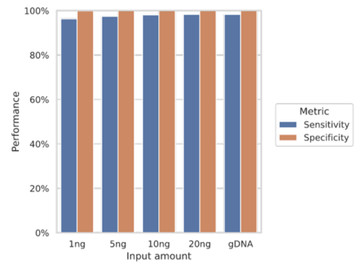
4A Duplication Rate



4B Genome Coverage



4C Sensitivity and Specificity



4D Read-level Methylation and Genetics

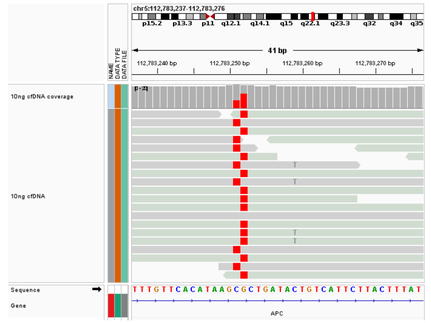


FIGURE 4: 5-LETTER SEQ PERFORMANCE and ALLELE-SPECIFIC METHYLATION DETECTION on CFDNA

A) **Duplication rate calculated by Picard MarkDuplicates**

B) **Proportion of the reference genome covered by at least 1 read**.

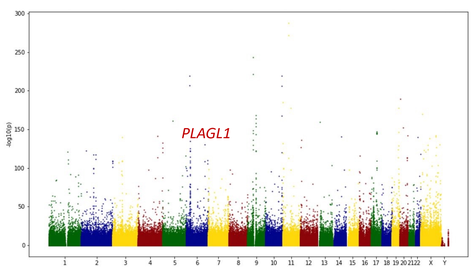
C) **Sensitivity and specificity on ground–truth spike-ins** described in figure 2.

D) **Integrated genomic view of region of *APC* gene** containing a C->T mutation in a minority of reads proximal to a hypermodified CpG.

CEGX SEQUENCING TECHNOLOGY PRODUCES PHASED EPIGENETIC AND GENETIC INFORMATION

CEGX sequencing technology generates phased discrimination of genetic and epigenetic letters on the same read. CEGX 5-Letter seq data demonstrates genome-wide allele-specific methylation and an example of A->G variant in the *PLAGL1* gene associated with *in cis* modified cytosine levels.

5A Genome-wide Allele-Specific Methylation



5B Phased Epigenetic and Genetic Information

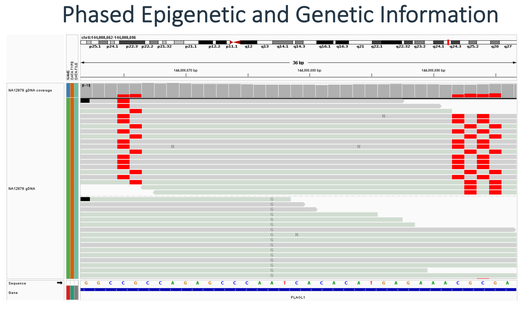


FIGURE 5: CEGX 5-LETTER SEQ GENOME-WIDE and READ SPECIFIC ALLELE SPECIFIC METHYLOME ANALYSIS

A) **Manhattan plot of allele-specific methylation in NA12878**. x-axis is chromosomal location and y-axis is –log10 p from fisher's exact test of association between genotype and in cis methylation levels.

B) **Integrated genomics viewer of 5-Letter sequence reads from NA12878 covering a region of *PLAG1* gene** identified with A->G heterozygous variant associated with *in cis* modified cytosine levels. Red boxes denote modified C calls which are exclusively found on reads also containing the A allele, whilst reads containing the G allele are unmodified at the same CpGs.

6-LETTER SEQ MAINTAINS HIGH ACCURACY

CEGX 6-Letter seq generates complete genetic information and discriminates between 5mC and 5hmC.

% Calls	pUC19 (C)	Lambda (mC)	Synthetic oligonucleotide (hmC)
C	99.880	3.45	1.64
mC	0.095	95.15	2.07
hmC	0.025	1.40	96.29

TABLE 1: CEGX 6-LETTER SEQ MAINTAINS GENETIC ACCURACY AND CALLING OF MODIFIED BASES WHILST ALSO DISCRIMINATING 5mC and 5hmC

Rate of calling the correct base at read level (row) on 3 different sources of ground truth (columns).

CONCLUSION

We describe a technology that delivers very high accuracy for both genetic and epigenetic sequencing by expanding the number of information states in next-generation-sequencing to sixteen. This allows direct, digital and phased discrimination of genetic and epigenetic letters on the same read, and error suppression.

The expanded information and enhanced accuracy afforded by the described technology enables:

- Full genetic and epigenetic information in a single workflow
- Phased combined genetic and epigenetic marks
- Decreased limit of detection for low frequency variant discovery
- Reduced consumption of sample and sequencing reagents, compared to multiple workflows for traditional NGS approaches.

Early access to the first product from this platform, CEGX 5-Letter seq is available, and CEGX 6-Letter seq is planned for Q4 2022.

cambridge epigenetix