Expanded information and enhanced accuracy of cell free DNA sequencing for early cancer detection

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

Detecting and treating cancer early will save lives. Minimally invasive blood testing, often called liquid biopsy, is a very attractive approach for early screening.

Liquid biopsy assays profile circulating tumour derived DNA (ctDNA) present in the cell free fraction of blood. Overcoming the technical challenge in sequencing and analysis of ctDNA, present in low quantities in a high background of normal DNA, requires enhanced approaches. The addition of epigenetic letters to the genetic sequence, increases sensitivity for early cancer detection and is informative of tissue of origin indicating the anatomical location of the tumour.[1,2] Existing DNA sequencing technologies are unable to simultaneously measure both genetic and epigenetic letters without sacrificing information content and accuracy. An alternative approach would be to perform multiple analyses on aliquots from the same sample. However, multiple analyses are not only expensive in terms of sample volume, time and reagent costs, but also yield sub-optimal information due to the inherent inaccuracy of data integration

Cambridge Epigenetix (CEGX) present 5-Letter seq. A technology that enhances the accuracy of cfDNA sequencing by expanding the information content of next-generation-sequencing from 4 states to 16. Furthermore, the information is phased, and interaction of genetics and epigenetic marks can be observed within single reads.

Materials & Methods

To demonstrate the technology's capabilities, DNA standards and human genomic DNAs with known genetic and epigenetic states, as well as cfDNA samples donated by individuals diagnosed with cancer were used to prepare sequencing libraries. These prepared libraries were sequenced in a standard Illumina (ILMN) NovaSeq sequencing run.

1 Single Workflow for Genetic and Epigenetic Information

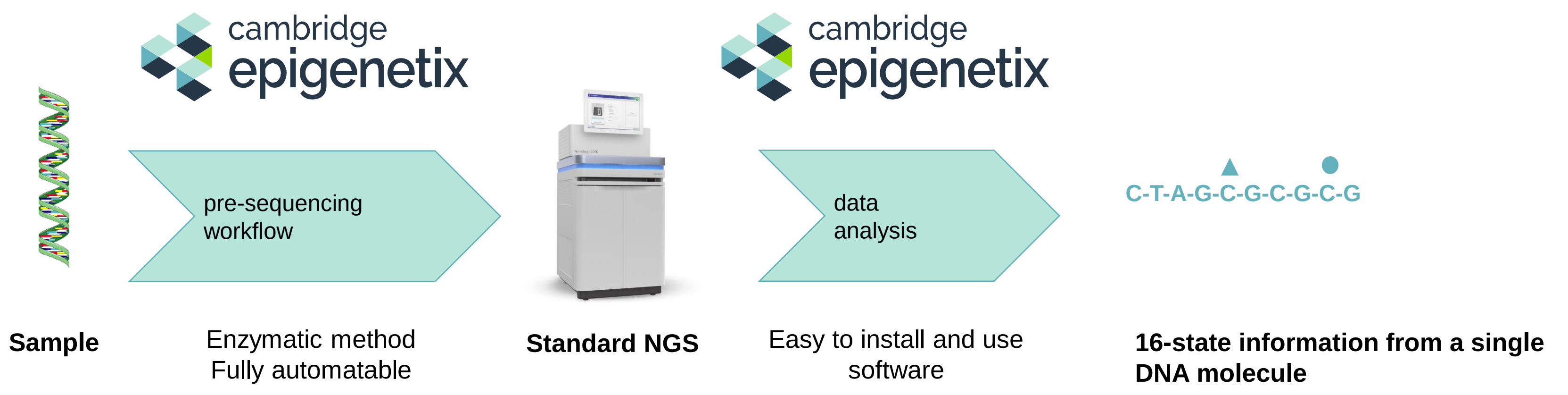
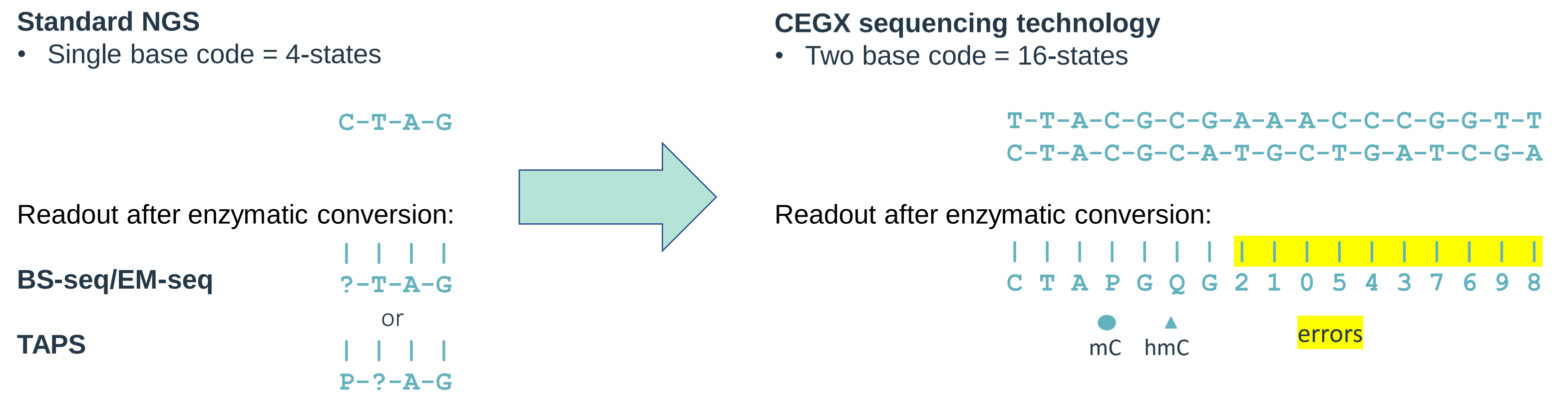


Figure 1. CEGX sequencing technology

CEGX technology is comprised of a single pre-sequencer workflow and post-sequencer software. This sequencer agnostic technology, initially optimised for ILMN sequencers without customisation, is presented in combination with a software pipeline that can be deployed on-premises and on all major cloud providers.

2 Genetic and epigenetic letters are digitally discriminated


Standard NGS
Single base code = 4-states
C-T-A-G
Readout after enzymatic conversion:
BS-seq/EM-seq
?-T-A-G
or
TAPS
P-?-A-G
CEGX sequencing technology
Two base code = 16-states
T-T-A-C-G-C-G-A-A-A-C-C-C-G-G-T-T
C-T-A-C-G-C-A-T-G-C-T-G-A-T-C-G-A
Readout after enzymatic conversion:
C T A P G Q G 2 1 0 5 4 3 7 6 9 8
mC
hmC
errors

BS-seq: Bisulfite sequencing
EM-seq: Enzymatic Methyl-seq

Figure 2. Schematics of 5-letter epigenetic sequencing protocol

Traditional genetic sequencing represents 4 information states and does not allow the identification of all letters of interest. Protocol changes, altering the information output, are still limited by only having four output states (BS-seq/EM-seq). Considering pairs of bases results in 2^4=16 possible states, allowing for the simultaneous detection of epigenetic and genetic information. The 16-state encoding also enables suppression of artefacts introduced during sample preparation and sequencing.

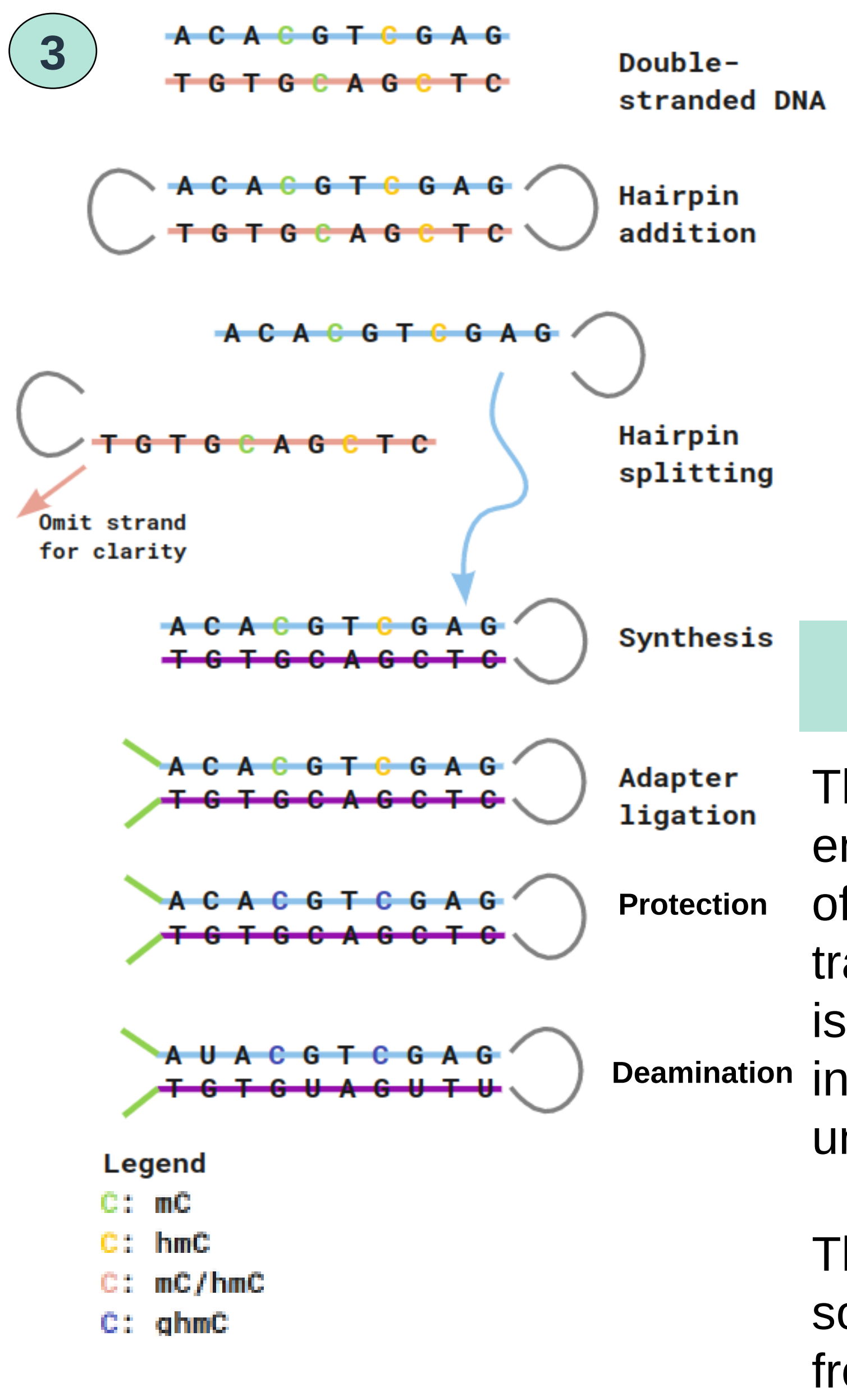


Figure 3. CEGX technology workflow

Hairpins are ligated to double-stranded DNA and the strands are separated. A copy strand is synthesised followed by short adapter ligation. Next, two enzymatic conversion steps are performed: 1) to protect modified cytosines (5-methylcytosine/5-hydroxymethylcytosine), and 2) to convert unmodified cytosine bases to uracil. Following library amplification, the final library is ready for paired-end sequencing. The expanded number of information states from 4 to 16 allows direct, digital and phased discrimination of genetic and epigenetic letters on the same read with suppression of PCR and sequencing errors.

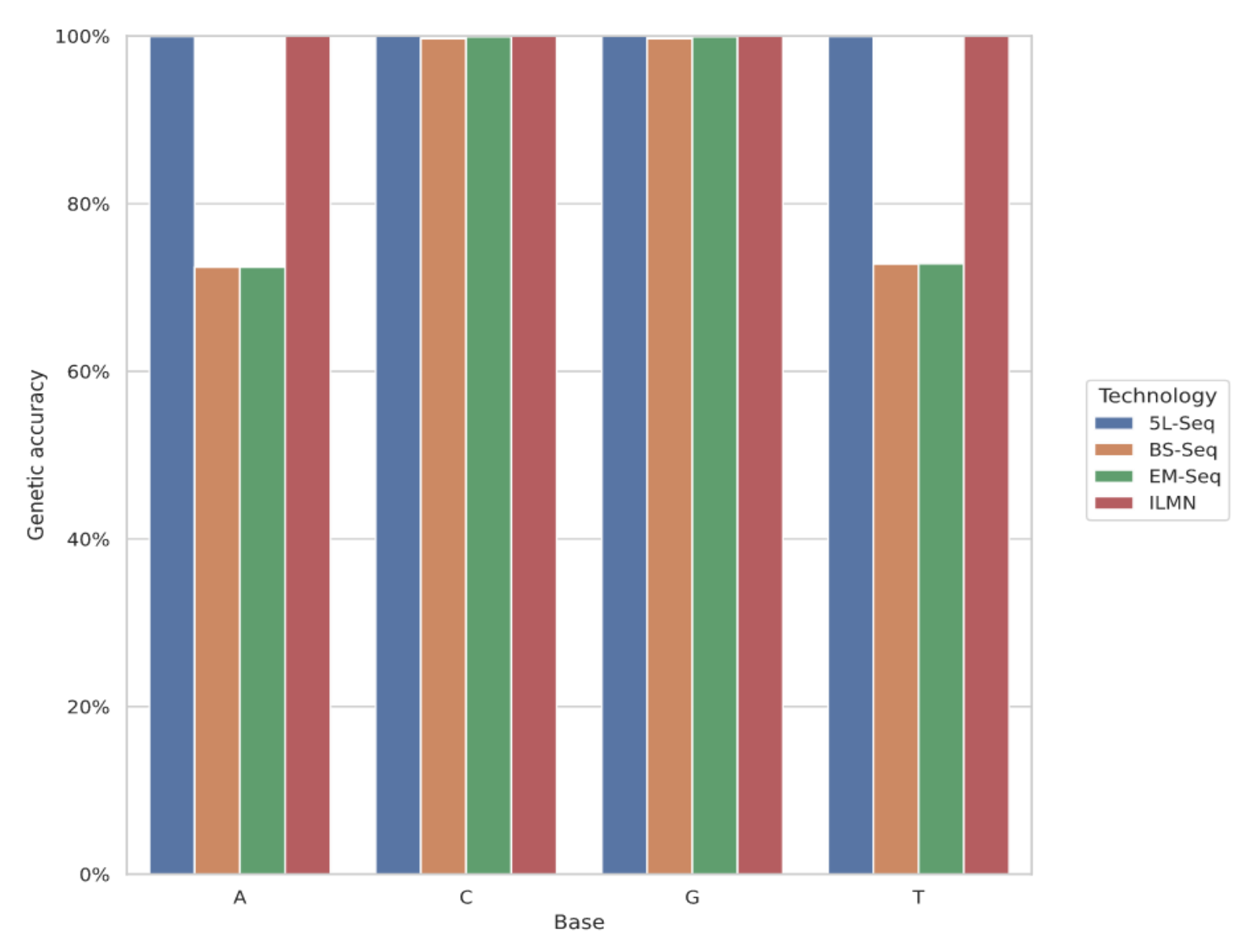
Results

5-Letter seq delivers best-in-class genetic and epigenetic accuracy.

Genetic accuracy, at read level, of CEGX 5-Letter seq was determined by standard whole genome sequencing (ILMN) of 'genome-in-a-bottle' sample (NA12878) and compared to BS-seq and EM-seq (Fig 4.). More accurate genetic information was achieved using CEGX technology than BS-seq and EM-seq because C>T mutations (the most common mutation type) are retained. 5-Letter seq results in accurate Phred scores of which 47% are above Q37, this being the highest attainable Phred score for ILMN.

Methylation accuracy was determined using ground-truth spike-ins. The comparisons demonstrate a higher accuracy of 5-Letter seq for detection of modified cytosine (modC) compared to BS-seq and EM-seq.

4 Genetic Accuracy



5 Methylation Accuracy

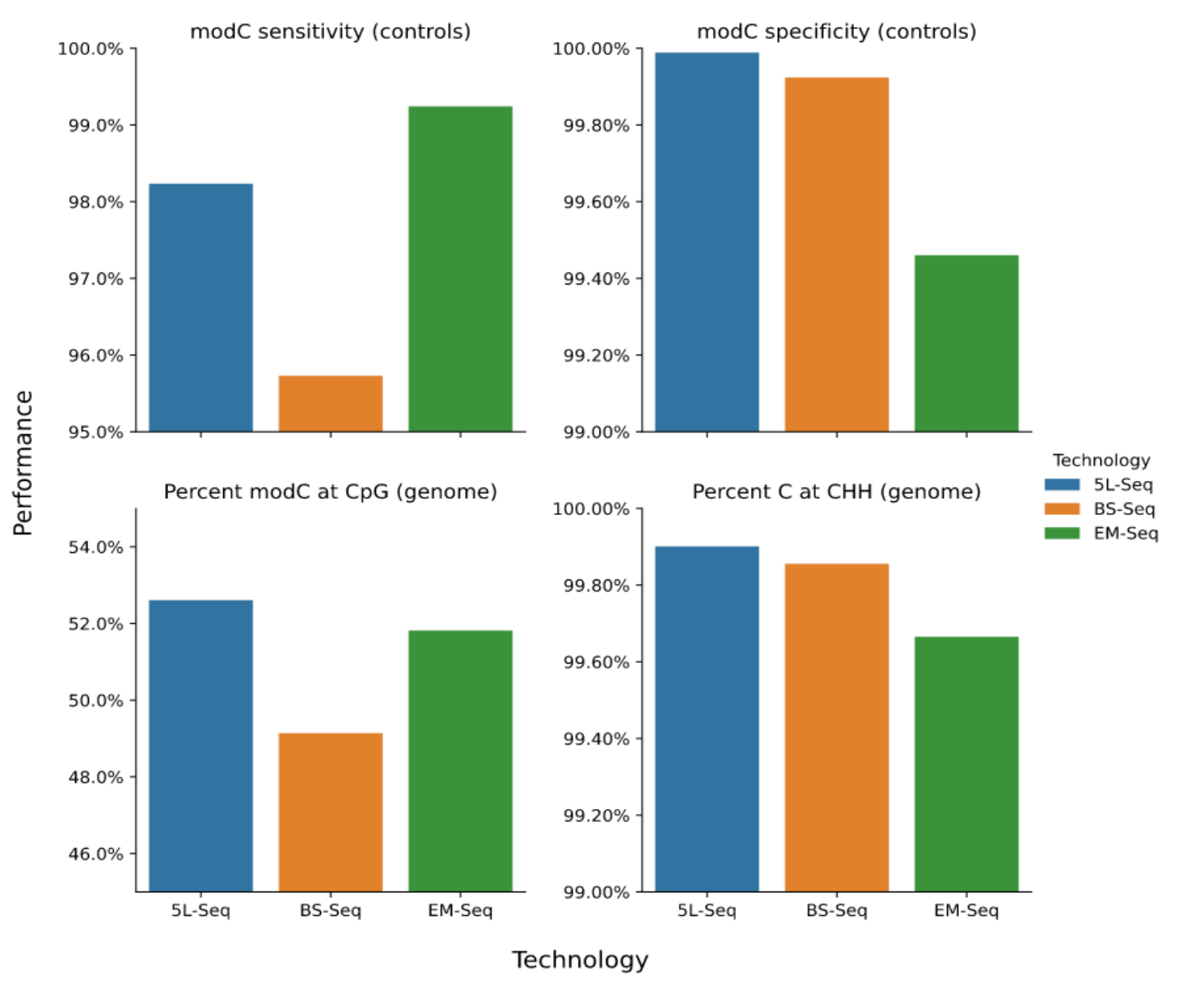


Figure 4. Genome-wide read level accuracy

- Based on sequencing NA12878 (80-100ng input) on a NovaSeq
- **Per base read-level genetic accuracy** for CEGX 5-Letter seq, BS-seq, EM-seq and ILMN, compared to gold-standard genome-in-bottle sequence. Only bases with Q-scores ≥ 25 were included

Figure 5. Accuracy of modified cytosine detection

5-Letter seq (blue) combined sensitivity and specificity measured on ground-truth controls is higher than either BS-seq (orange) or EM-seq (green). 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 in 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

5-Letter seq is compatible with liquid biopsy.

5-Letter seq was performed on cell free DNA (cfDNA) from a patient with stage 3 colorectal cancer (CRC) ranging from 1ng-20ng input (Fig. 6)

- Data shows compatibility with the cfDNA quantity typically available from a standard blood draw (Fig.6)
- Fig. 8. A C->T mutation within the *APC* gene, present in a minority of reads, proximal to a hypermodified CpG.

6 Genome Coverage

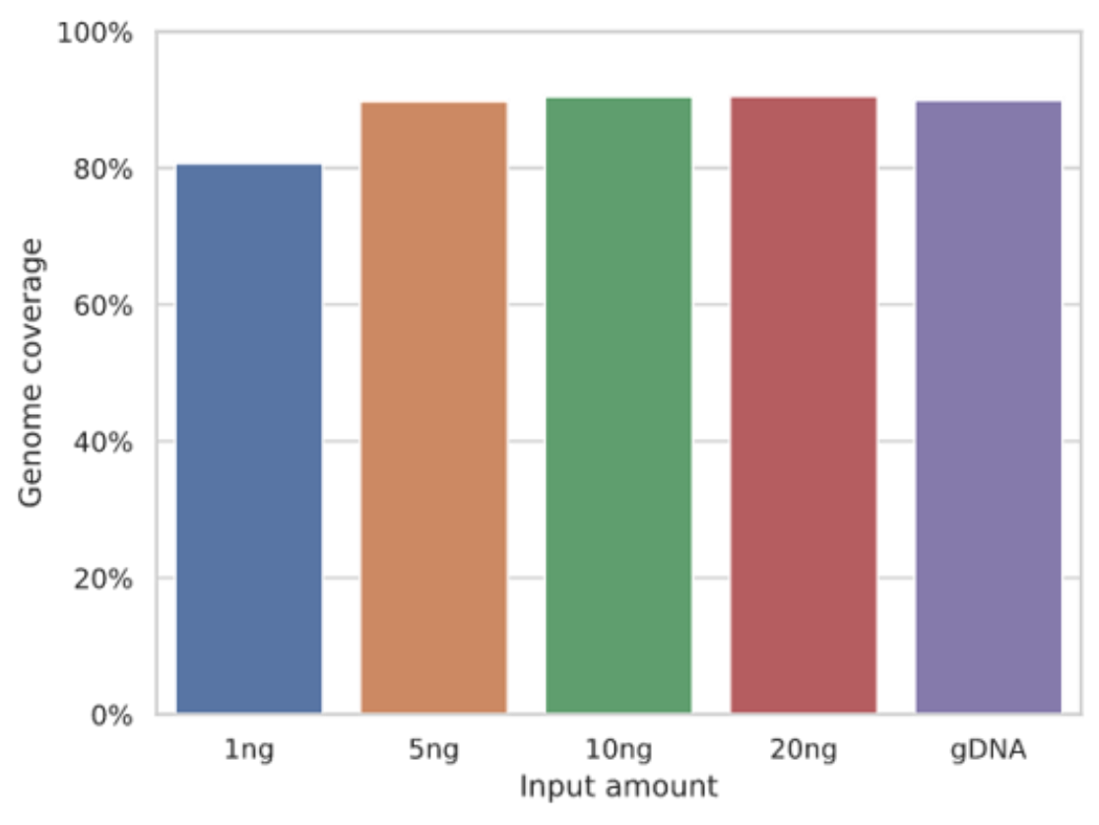


Figure. 6. **Proportion of the reference genome covered by at least 1 read**.

7 Sensitivity and Specificity

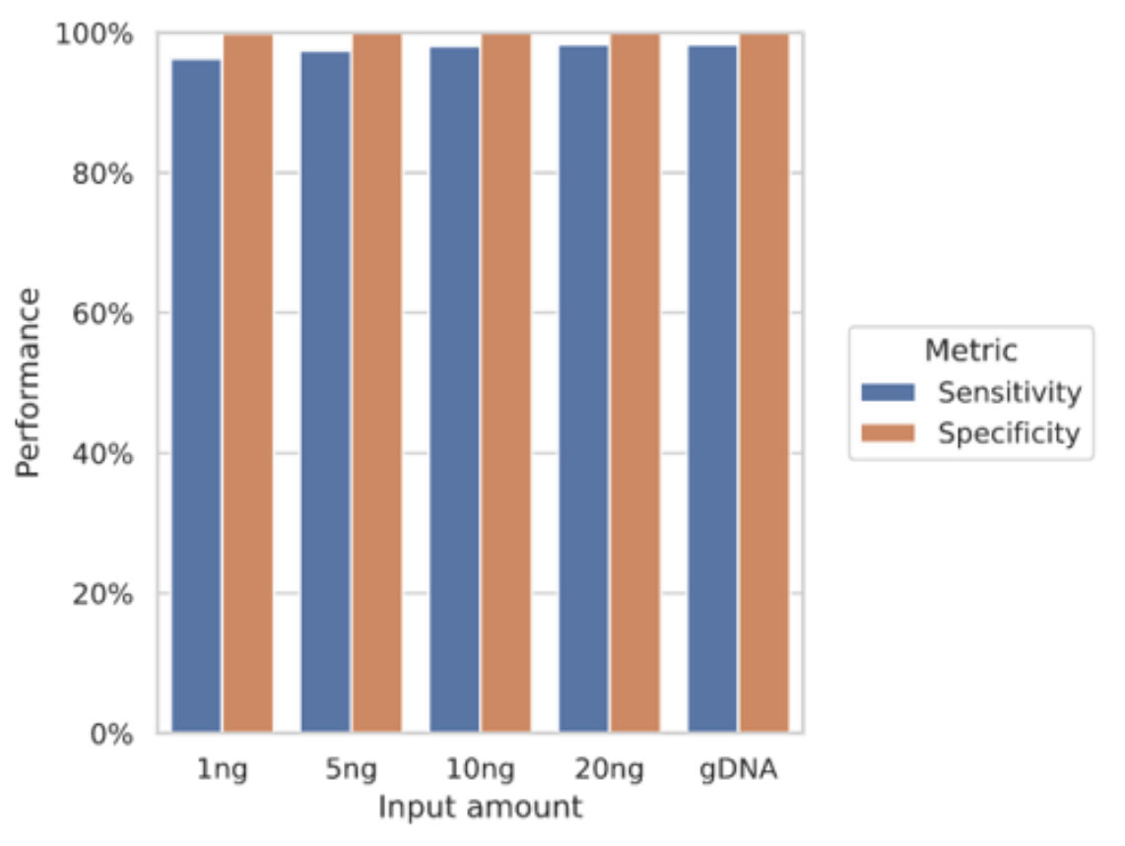


Figure. 7. Sensitivity and specificity on ground–truth spike-ins described in figure 5.

8 Read-level Methylation and Genetics

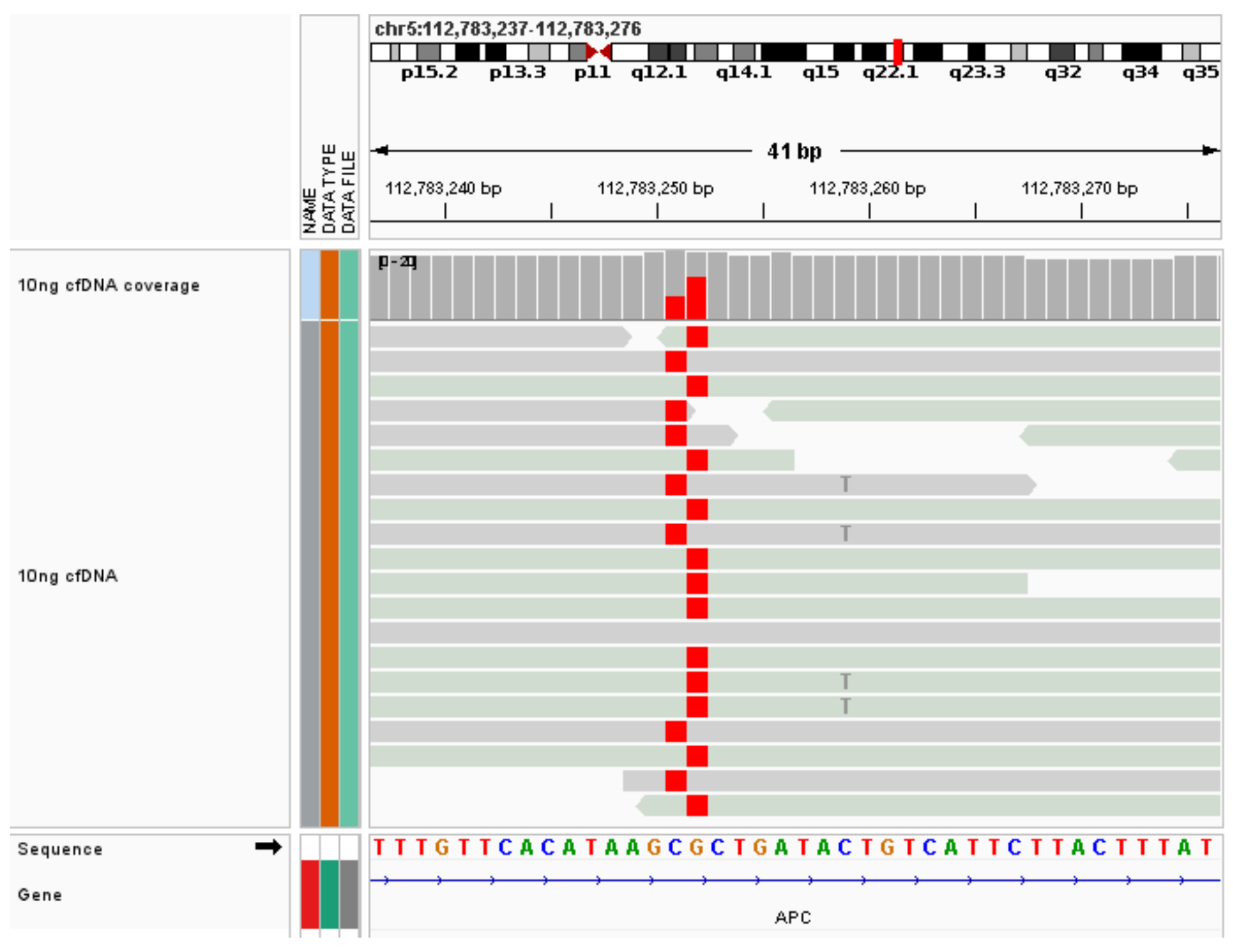


Figure. 8. Integrative Genomics Viewer (IGV) image of a region of *APC* gene showing a C->T mutation in a minority of reads proximal to a hypermodified CpG.

Conclusions

The expanded information and enhanced accuracy afforded by the described technology enables simultaneous phased genetic and epigenetic data to be produced from nanograms of cfDNA and a single workflow. Accurately combining genomics and epigenomics will translate to higher sensitivity for detection of new or recurrent cancer[3]. That this information is provided in the same read and from the same molecule is important, as studying the interaction of genetics and epigenetic marks is necessary for a more complete understanding of oncogenesis and tumour biology[4,5].

The technology comprises a pre-sequencing molecular biology kit and post-sequencing software. It is compatible with existing sequencers and compute set-ups. The first product from this platform, 5-Letter seq, is now available to early-access customers; with 6-Letter seq available at the end of 2022.

References:

1. *Crosby, D. et al Early detection of cancer. 11 (2022)*
2. *Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science 372, eaaw3616 (2021).*
3. *Kim, S.-T. et al. Abstract 916: Combined genomic and epigenomic assessment of cell-free circulating tumor DNA (ctDNA) improves assay sensitivity in early-stage colorectal cancer (CRC). 916-916, doi:10.1158/1538-7445.sabcs18-916 (2019).*
4. *Feinberg, et al The Epigenetic Progenitor Origin of Human Cancer*
5. *Alonso-Curbelo (2021) A gene-environment-induced epigenetic program initiates tumorigenesis*