

AI-Driven Tumor Detection from cfDNA Methylation Using the Illumina 5-Base Sequencing Kit

Date: 12 March 2026

1. Introduction

Cell-free DNA (cfDNA) methylation profiling has emerged as a powerful approach for non-invasive cancer detection and monitoring. The ability to detect tumor-derived cfDNA (ctDNA) in plasma relies on identifying cancer-specific methylation patterns that differ from those of normal tissues.

The Illumina 5-Base Sequencing Kit is a reagent kit enabling direct detection of 5-methylcytosine (5mC) during short-read sequencing, providing base-level methylation information. This technical note describes its application to a cohort of plasma cfDNA samples from cancer patients and healthy donors. Tumor scores computed using Aíma's AI-based ctDNA detection model from 5-Base Illumina sequencing show a strong correlation with those obtained from Oxford Nanopore Technologies (ONT) sequencing.

2. Experimental Design

2.1 Sample Cohort

A total of 8 anonymised cfDNA plasma samples were processed in this study (Samples 1–8), encompassing three cancer tissue types and a healthy donor control group:

- Breast tumor: 1 sample (Sample 1)
- Colon tumor: 3 samples (Samples 2, 3, 4)
- Lung tumor: 2 samples (Samples 7, 8)
- Healthy donors: 2 samples (Samples 5, 6)

Each sample was sequenced using both the Illumina 5-Base Kit (short-read) and ONT (long-read), with ONT providing an independent tumor score used as reference for comparison.

2.2 Sequencing Platforms

ONT long-read sequencing was performed using **two P4 flow cells**. Short-read sequencing with the Illumina 5-Base Kit was carried out on a **NextSeq 2000** instrument.

2.3 Workflow Overview

cfDNA was extracted from plasma samples and processed according to the Illumina 5-Base Sequencing Kit library preparation protocol. Raw reads were quality-filtered, adapter-trimmed, and aligned to the human reference genome (GRCh38). Methylation calls were extracted at CpG sites and used as input for tumor score estimation. The same samples were independently sequenced on the ONT platform using two Promethion flow cells to generate reference tumor score values.

3. Sequencing Quality Metrics

Table 1 summarises the per-sample sequencing statistics generated by the Illumina 5-Base Kit on the NextSeq 2000, including total reads, mapped reads, sequencing depth, and genomic coverage.

Sample	Tissue of origin	Total reads (M)	Mapped reads (M)	Depth (×)	Coverage (% of the genome covered)
Sample 1	Breast	1,099.00	1,091	19.70	90.66
Sample 2	Colon	576.00	566	12.21	93.87
Sample 3	Colon	943.00	915	15.56	94.04
Sample 4	Colon	918.00	909	14.92	93.88
Sample 5	Healthy	572.00	493	10.89	94.27
Sample 6	Healthy	555.00	534	11.25	94.20
Sample 7	Lung	828.00	799	13.11	94.07
Sample 8	Lung	1,976.00	1,959	42.29	94.19

Table 1. Per-sample sequencing metrics generated with the Illumina 5-Base Kit on the NextSeq 2000.

Figure 1 illustrates the total and mapped read counts across all 8 samples. Total read counts ranged from approximately 555 million (Sample 6) to over 1,976 million (Sample 8). Mapped read counts were broadly comparable across samples, reflecting consistent alignment efficiency. Genomic coverage was uniform across most samples (~93–94%), with Sample 1 (Breast) at 90.66%.

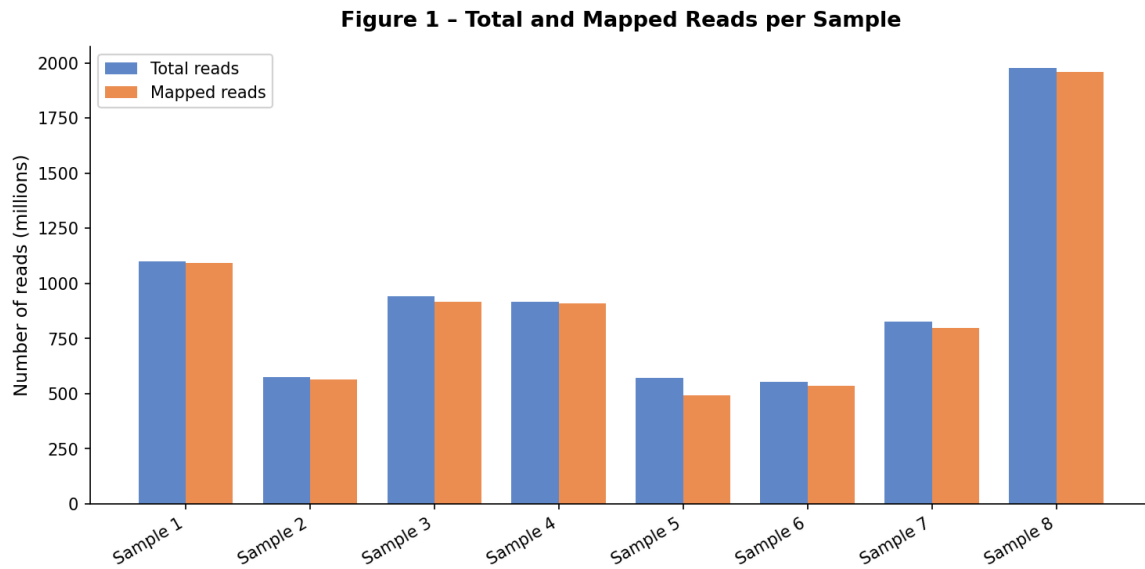


Figure 1. Total reads (blue) and mapped reads (orange) per sample in millions (M).

Sequencing depth was robust across the cohort (Figure 2), ranging from 10.89× (Sample 5) to 42.29× (Sample 8).

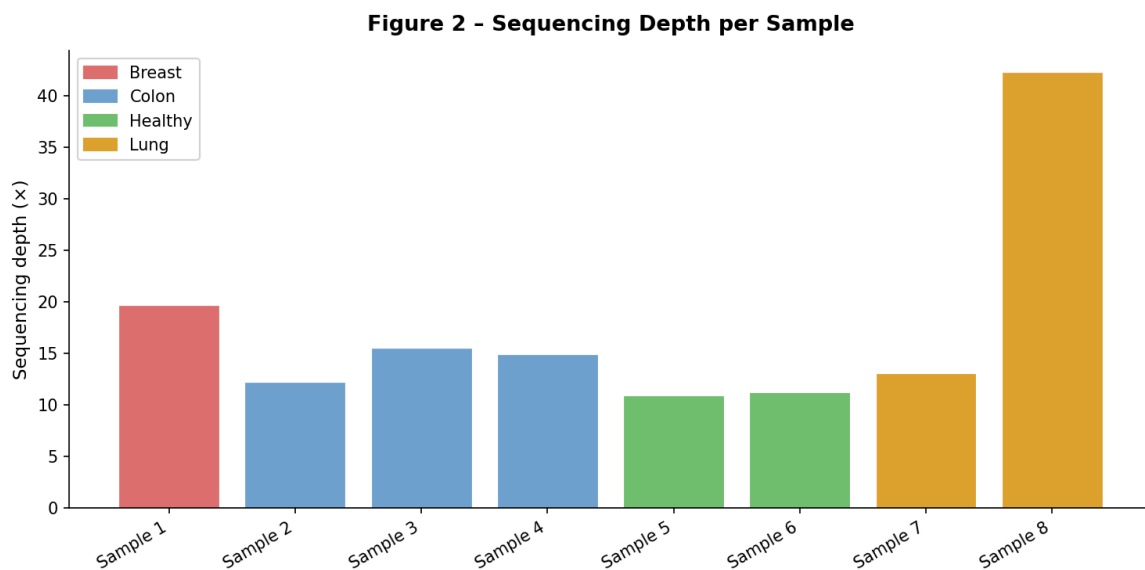


Figure 2. Per-sample sequencing depth (x). Colour coding: Breast (red), Colon (blue), Healthy (green), Lung (orange).

4. Tumor Score provided by Aíma's AI-based ctDNA detection model

Tumor score was estimated from the Illumina 5-Base methylation data and compared to the ONT-derived values (Table 2). The results demonstrate a clear and consistent pattern: the 5-Base Illumina tumor score increases proportionally with the ONT tumor score, confirming

that the Illumina 5-Base Kit faithfully captures the level of tumor-derived signal present in the sample. Notably, the two healthy donor samples (Samples 5 and 6) yielded the lowest tumor scores across the entire cohort with both methods, which demonstrates the ability of the assay to discriminate tumor-bearing from healthy samples. Table 2 is ordered by 5-Base Illumina tumor score in ascending order.

Sample	Tissue of origin	ONT tumor score	5-Base Illumina tumor score
Sample 6	Healthy	0	1.70
Sample 5	Healthy	0	1.74
Sample 3	Colon	0	2.185
Sample 4	Colon	0.2103	2.185
Sample 2	Colon	1.667	2.281
Sample 7	Lung	1.222	2.457
Sample 8	Lung	20.56	11.60
Sample 1	Breast	20.56	15.73

Table 2. Comparison of tumor score estimates between ONT and the Illumina 5-Base Kit (NextSeq 2000), ordered by ascending 5-Base Illumina tumor score.

Figure 3 presents a scatter plot of the ONT and 5-Base Illumina tumor scores. The two methods are very strongly correlated, with a Pearson correlation coefficient of $r = 0.978$, demonstrating excellent concordance across the full range of tumor scores observed in this cohort. Samples with high tumor burden (Samples 1 and 8) cluster at the upper end of both axes, while healthy donors (Samples 5 and 6) and low-score samples are consistently positioned at the lower end, further supporting the clinical relevance of the 5-Base Illumina tumor score. Additional samples will further support accurate definition of the threshold to discriminate positive from negative thresholds.

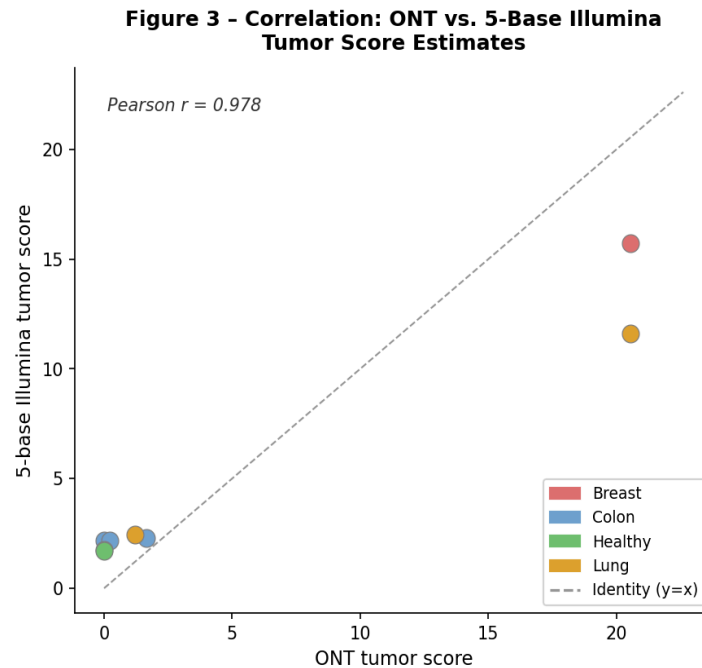


Figure 3. Scatter plot of ONT (x-axis) vs. 5-Base Illumina (y-axis) tumor score estimates. The dashed line represents the identity line ($y = x$). Pearson $r = 0.978$. Colour coding indicates tissue of origin.

5. Discussion

The Illumina 5-Base Sequencing Kit, run on the NextSeq 2000, generated high-quality methylation data across all 8 cfDNA samples, with consistent mapping rates and coverage across the cohort. Tumor scores computed using Aíma's AI-based ctDNA detection model from 5-Base Illumina sequencing show a very strong positive correlation with the ONT reference values (Pearson $r = 0.978$), demonstrating the reliability and quantitative accuracy of the approach across a range of tissue types and tumor burden levels.

A particularly encouraging finding is that the 5-Base Illumina tumor score is positively associated with the level of tumor-derived signal: the higher the tumor content in the sample, the higher the score assigned by the Illumina assay. Healthy donor samples consistently received the lowest scores in the cohort, reinforcing the specificity of the assay for tumor-derived methylation signal.

6. Conclusions

This study demonstrates that the Illumina 5-Base Sequencing Kit, sequenced on the NextSeq 2000, delivers comprehensive and high-quality methylation profiles from cfDNA plasma samples. Tumor scores computed using Aíma's AI-based ctDNA detection model show very strong concordance with those obtained from ONT sequencing (Pearson $r = 0.978$), validating the quantitative performance of the approach. The assay successfully ranked healthy donors as the lowest-scoring samples, highlighting its potential for distinguishing tumor-bearing individuals from healthy controls. These findings position the Illumina 5-Base Kit as a highly promising technology for methylation-based liquid biopsy applications when combined with Aíma's AI-based analysis framework.