JOURNAL OF
EMERGING INVESTIGATORS

Article

Identifying 5-hydroxymethylcytosine as a potential cancer biomarker using FFPE DNA samples

Allison Li[1], Qing Dai[2]

[1] The University of Chicago Laboratory High School. Chicago, Illinois

[2] Department of Chemistry, The University of Chicago, Chicago, Illinois

SUMMARY

Head and neck cancer (HNC) is the seventh most common cancer worldwide and identifying biomarkers for its diagnosis, prognosis, and treatment success monitoring is crucial. Recent studies suggested that 5-hydroxymethylcytosine (5hmC), a modified DNA base that can impact gene expressions, has great potential as a biomarker for the diagnosis of various cancers. However, whether 5hmC can be used as a biomarker for HNC detection remains unexplored. In contrast to fresh HNC tissues from patients, formalin-fixed and paraffin-embedded (FFPE) samples are much more readily available for biomarker seeking. However, FFPE samples usually have severe DNA damage, leading to a large percentage of ssDNA fragments and making the sample incompatible with the current technology for 5hmC profiling. We hypothesized that using a new approach for 5hmC profiling that is compatible with ssDNA fragments would allow us to identify the potential 5hmC biomarker for HNC diagnosis in FFPE samples by comparing the differences in 5hmC distribution between tumor and adjacent normal tissues. Here, we report the use of an improved CMS-seq. method to sequence FFPE samples from HNC tumors and their adjacent normal tissues to generate pairs of genomic data from four different patients. After sequencing and data analysis, we identified 339 genes with differentiable 5hmC levels (p-value <0.05). Among them, three genes (PRKD2, HADHA, and AIPL1) have p-adjusted values less than 0.05, suggesting that the distinct 5hmC pattern in these three genes has promising potential as biomarkers for HNC diagnosis.

INTRODUCTION

Head and neck cancer (HNC) is the seventh most common cancer worldwide, comprising ~4.5% of cancer diagnoses and 4.6% of cancer deaths (1). The most common subtype, head and neck squamous cell carcinoma (HNSCC), accounts for about 90% of HNC cases (2). HNSCC is one of the most aggressive cancers, and more than half of the diagnosed patients have a survival rate of less than five years (3,4). The treatment of HNSCC is complicated by its heterogeneity, or variation in tumor characteristics among patients, often leading to locoregional recurrence, where cancer returns to the original tumor site or nearby regions, or distant metastasis, the spread of cancer to other parts of the body (2). Current treatments, including surgical removal of cancer, radiotherapy, chemotherapy, and targeted drugs like cetuximab, are limited in their effectiveness due to the unique and diverse nature of HNCSS (5). This underscores the urgent need for early diagnosis, especially in light of the projected 30% increase in HNSCC cases by 2030 (1).

Epigenetic modifications in the human genome, including histone modifications and nucleic acid methylations, are critical in regulating gene expression by modulating transcription factor binding (6). These modifications are particularly relevant in HNC, where they play a significant role in carcinogenesis and cancer development (5). For example, DNA methyltransferase catalyzes the formation of 5-methylcytosine (5mC) within the DNA, occurring predominantly at CpG sites, where a cytosine base is followed by a guanine base in the 5' to 3' direction (3). 5mC can undergo active demethylation catalyzed by Ten-eleven translocation (TET) proteins through stepwise oxidation to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (7,8). This 5mC active demethylation pathway is crucial since it influences gene expression by controlling chromatin accessibility and transcription repression (9). Unlike 5fC and 5caC, 5hmC remains relatively unchanged through the cell cycle, a necessary feature of inheritable epigenetic markers (10). 5hmC content levels vary depending on tissue type, from 0.40-0.65% (e.g. in brain, rectum, and colon) to 0.05-0.06% (e.g. in heart, breast, and placenta tissues) (11). In many types of cancer, 5hmC has been found at decreased levels (12). Though 5hmC is rare and less studied than 5mC, it may be better at reflecting gene expressivity through its correlation with active transcription unlike 5mC, which is generally associated with gene repression (13). 5hmC is often located at enhancer regions or active gene bodies and makes 5hmC a potential indicator of not just the state of transcription activity but also a measure of change in such activities (13). At the same time, 5hmC is a more malleable modification than 5mC because 5hmC is an intermediate in the active DNA demethylation process, meaning 5hmC can be more readily reversed or altered by enzymes such as TET (Ten-Eleven Translocation) proteins. This property makes 5hmC a dynamic marker that can reflect the state of biological processes more immediately and suggests that 5hmC has great potential as a novel biomarker for disease diagnosis and prognosis.

In recent years, promising research has been conducted on the use of 5mC and 5hmC as biomarkers for early cancer diagnoses (14). For example, two 5mC biomarkers for colon cancer diagnosis have been approved by the FDA, and products such as Cologuard have been used for the annual screening of stool samples for people older than 50 years old (15). In head and neck cancer, 5hmC levels have been found to significantly correlate with tumor stages and recurrence,

https://doi.org/10.59720/24-060

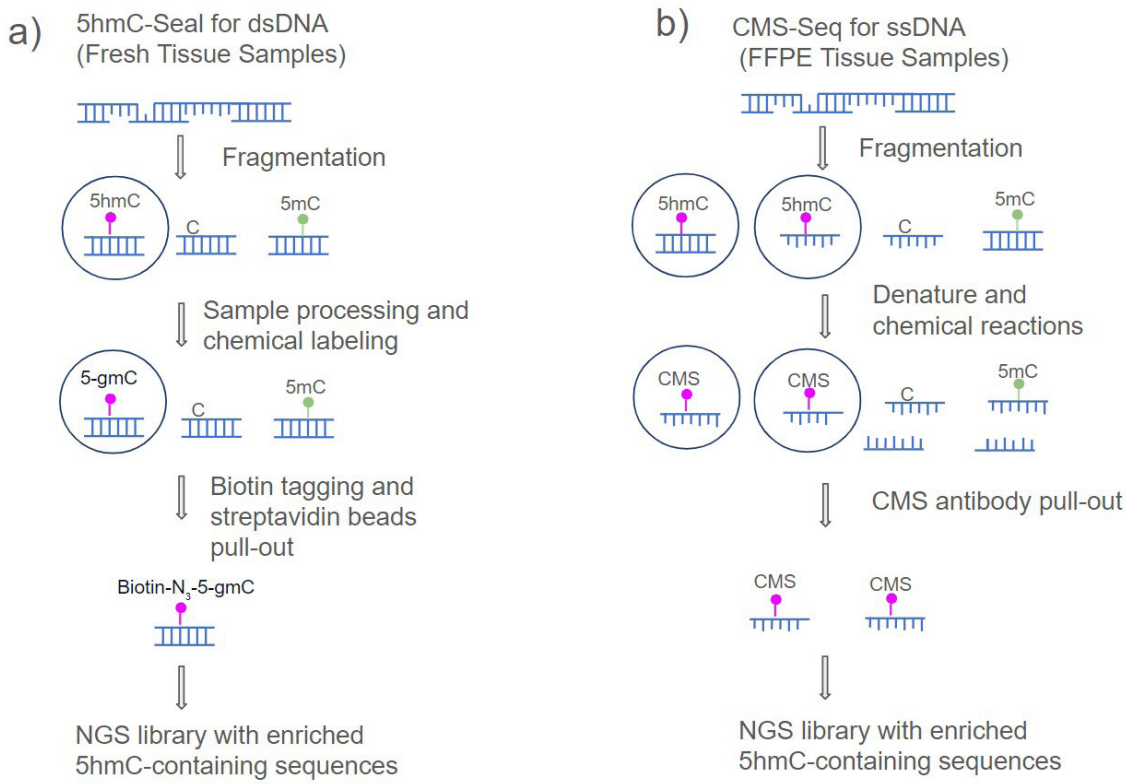

Figure 1: Comparison of the workflow of 5hmC-Seal and CMS-seq. FFPE DNA was first fragmented to ~200 bp length. In the 5hmC-Seal method (left), 5hmC is labeled with a glucose moiety using T4 β-glucosyltransferase, which allows selective pull-down of 5hmC-modified dsDNA fragments, as the method only works for double-stranded DNA. The CMS-seq method (right) works better for FFPE samples since CMS-seq works for both dsDNA and ssDNA fragments. 5hmC on dsDNA and ssDNA fragments were converted into CMS and all 5hmC-containing fragments were then pulled down by an anti-CMS antibody. In both methods, NGS libraries were constructed from the enriched 5hmC-containing sequences.

and low levels of 5hmC are associated with poorer survival (4). However, whether 5hmC can serve as a biomarker for HNC diagnosis has not been explored. One of the key reasons is that fresh tumor tissues with intact double-stranded DNA (dsDNA) are usually hard to obtain in large numbers from patients. On the other hand, formalin-fixed and paraffin-embedded (FFPE) samples are readily available and have been indispensable in preserving biopsy specimens for years in cancer diagnosis and research (16). The longevity of FFPE samples and cost-effective storage have led to an expansive FFPE sample collection, making up the vast majority of all available biological samples (16). Therefore, FFPE samples are a great resource for seeking biomarkers. However, they have been underutilized in genomic sequencing studies compared to fresh-frozen tissue samples (17,18). FFPE samples contain DNA of inferior quality because formalin fixation causes hydrolytic damage, fragmentation, and cross-linkages, leading to the formation of single-stranded DNA (ssDNA) with percentages as high as 60% (19). These ssDNA fragments are not compatible with current 5hmC profiling methods such as 5hmC-seal, which only works for dsDNA fragments (**Figure 1)** (20). Although very powerful, the 5hmC-Seal method requires dsDNA since the β-GT-catalyzed glycosylation reaction only works for dsDNA, but not ssDNA. Other methods, such as TET-assisted bisulfite sequencing and oxidative bisulfite sequencing, have drawbacks in 5hmC enrichment because they require significant sequencing depth and rely on harsh DNA treatments, such as bisulfite conversion (21,22). These treatments degrade DNA, potentially leading to incomplete conversion and biases against GC-rich regions, which lowers sequencing accuracy and makes further data analysis less reliable (22). Traditional cytosine-5-methylenesulfonate sequencing (CMS-seq) uses conventional bisulfite sequencing (BS) treatment to convert 5hmC to cytosine-5-methylenesulfonate (CMS) and then uses anti-CMS antibody for enrichment and sequencing (23). Despite its success, this method still suffers two limitations: (i) Conventional BS treatment causes an undesirable base conversion from C to deoxyuridine (dU), which reduces complexity and causes mapping issues; and (ii) conventional BS treatment also causes severe DNA damage, requiring high input of DNA (500 to 1000 ng).

To address this incompatibility, we hypothesized that a new approach for 5hmC profiling making use of ssDNA fragments would allow us to use FFPE samples to investigate the potential using 5hmC as a biomarker for HNC diagnosis

https://doi.org/10.59720/24-060

by comparing the differences in 5hmC distribution between tumor and adjacent normal tissues. Here, we developed an improved CMS-seq method for 5hmC enrichment and library construction and demonstrated that it can be used on FFPE DNA samples extracted from cancer and adjacent normal tissues. We improved CMS-seq by using a new BS recipe at neutral conditions, which can overcome these two limitations and successfully construct libraries from 50 ng of DNA extracted from FFPE tissues. We identified three genes showing distinct 5hmC distribution patterns, suggesting that they have promising potential as biomarkers for HNC diagnosis.

RESULTS

Improved CMS-seq method

We initially focused on optimizing a newly developed 5hmC profiling method termed CMS-seq, which works for both dsDNA and ssDNA, to improve library construction efficiency (**Figure 1**). Our developments help achieved 200-fold enrichment of CMS-containing fragments and were compatible with both dsDNA and ssDNA (**Figure 1)**. These improvements resulted in higher specificity, more reliable 5hmC profiling, and better overall data quality without the serious DNA damage or undesired C-to-U conversion observed in earlier methods (**Figure 1**).

Data analysis of CMS-seq libraries

We established a pipeline for data analysis and identified 14,606 genes for downstream analysis, selecting only genes with more than 50 total reads in at least one sample to ensure statistical robustness (**Figure 2**). Our confidence in the analysis increased as the differences in 5hmC-modification levels between tumor and normal tissue genes widened. Notably, tumor tissues have previously been shown to more often exhibit downregulation in 5hmC compared to normal ones (24). Our analysis identified 261 genes with tumor tissues containing less than half of the 5hmC content level compared to healthy tissues. On the other hand, only 78 genes showed higher 5hmC content. In total, 339 genes were identified as both statistically robust (p-value < 0.05) and having more

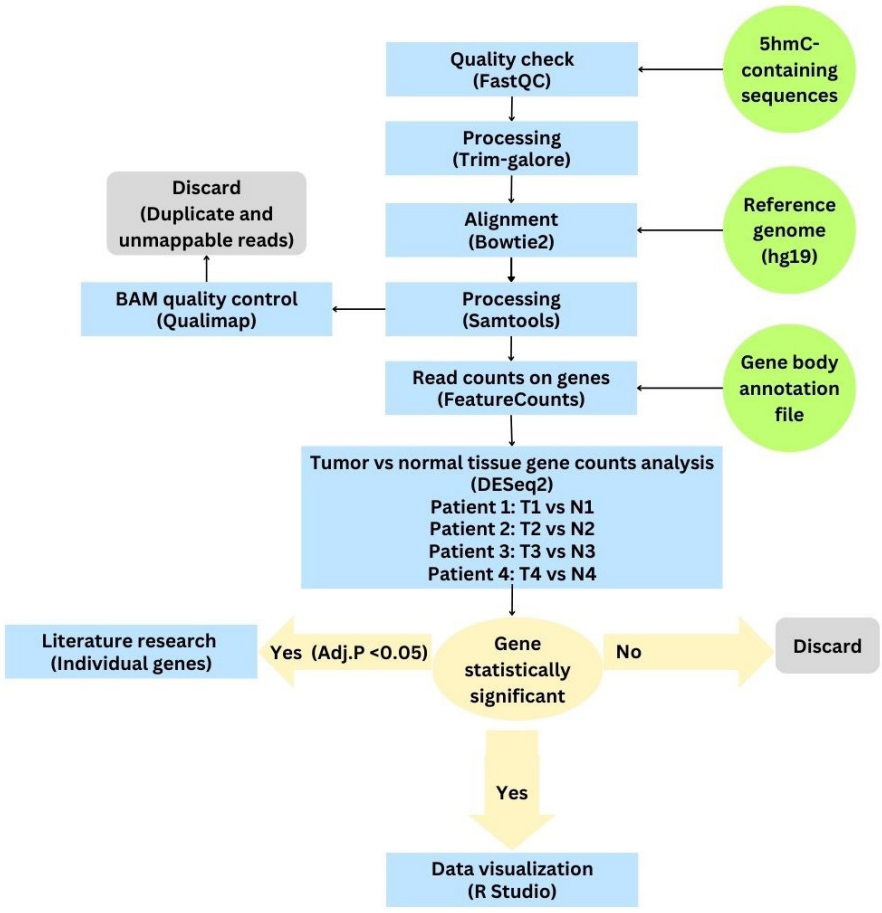

Figure 2: Data analysis workflow. The flowchart depicts the steps from processing the Next Generation Sequencing data to analysis and visualization in R Studio. After a quality check (FastQC) and trimming (Trim-galore), reads are aligned to the hg19 reference genome (Bowtie2), followed by processing (Samtools) and counting of gene reads (FeatureCounts). Statistically significant genes (p-adjusted value < 0.05) are identified using DESeq2, and non-significant genes are discarded. Significant genes undergo further literature research and data visualization. The reference genome was obtained from the National Center for Biotechnology Information publicly available database. The gene body annotation file was generated using R from the reference genome. Points labeled with the starting character "N" are plotted from normal samples and "T" from tumor samples. Points labeled with the same ending number are from samples from the same individual.

https://doi.org/10.59720/24-060

Figure 3: Volcano plot illustrating the 5hmC-modification differences in genes between the HNC tumors and adjacent normal tissue DNA. The two vertical green lines distinguish genes that have at least a 2-fold change in 5hmC concentration. Genes to the left of the leftmost green line have over a 2-fold decrease in 5hmC concentration between tumor tissues and normal tissues. Conversely, genes to the right of the rightmost green line have over a 2-fold increase in 5hmC concentration in tumor tissue as compared to normal tissue. The y-axis represents the -log10 of the probability value (p-value) of each gene, where a p-value under 0.05 is considered statistically relevant and sectioned above the blue line. Genes with p-adjusted values less than 0.05 using Benjamini-Hochberg procedure were colored in red to distinguish them as the most statistically accurate data in this set.

than a two-fold change in 5hmC levels amongst sample pairs (Figure 3). Additionally, we employed the *p*-adjusted value to enhance the statistical robustness of our findings. After adjusting for false positives using a *p*-adjusted value threshold of 0.05, we pinpointed three genes of particular interest: *PRKD2*, *HADHA*, and *AIPL1* (**Figure 3**).

Principal Component Analysis (PCA) is a statistical method that simplifies the complexity of a data set by transforming the original set of variables into a reduced number of principal components (25). The first principal component captures the largest variance in the dataset, and the second principal component accounts for the next largest variance, indicated by the x- and y-axis, respectively. The resulting plot allows us to gauge the similarity between samples based on the proximity of their respective points. In our analysis, using all 14,606 genes did not reveal a clear distinction between tumor and adjacent normal tissues (**Figure 4a**). However, a notable variation and separation were exhibited when only the 339 statistically significant genes were analyzed through PCA (**Figure 4b**). Without excluding the unique characteristics of each sample, PCA analysis shows that distinct 5hmC features can separate tumor tissues from adjacent normal ones, predominantly along the second principal component (**Figure 4**).

To highlight differences between normal and tumor tissues, we normalized the 5hmC levels for each gene using z-scores. Genes with low 5hmC levels are represented by lower z-scores, while higher z-scores indicate genes with elevated 5hmC modification. This normalization allowed us to clearly identify patterns of 5hmC modification across samples (**Figure 5**). There is a clear contrast in 5hmC levels between tumor and normal tissue DNA (**Figure 5**). Tumor tissue DNA generally exhibited significantly lower 5hmC levels (**Figure 5**). Intriguingly, one pair of tumor and normal samples from a patient displayed an unusually high level of 5hmC across multiple genes compared to the other three pairs of samples (**Figure 5**). The presence of the outlier pair suggests that this individual might possess unique circumstances resulting in a more saturated 5hmC landscape than identified in the other three pairs.

Candidate genes

PRKD2 is a gene in humans that encodes the Protein Kinase D2 (PKD2) enzyme and plays a critical role in various cellular processes, including cell proliferation, survival, migration, and the formation of new blood vessels (26). Our results indicated that *PRKD2* exhibited lower levels of 5hmC in tumor tissues than normal ones (**Figure 3**). Using Gene Expression Profiling Interactive Analysis (GEPIA 2), we mapped the expression levels of *PRKD2* in 519 HNC and 44 healthy patient samples from publicly available datasets: The Cancer Genome Atlas Program (TCGA) and The Genotype-Tissue Expression (GTEx), respectively (27). We found that the expression of the PRKD2 gene was much higher in HNC patients than in healthy controls (**Figure 6a**). ANOVA differential method was used for tumor versus paired normal sample. Although there was not a statistically significant difference, the mean transcription per million (TPM) rate in tumor samples suggested a higher gene expression trend of 43 while it was only 24 in normal samples. As prior studies suggested that low levels of 5hmC correlate with decreased gene expression, we concluded that a more complicated relationship exists between 5hmC transcription regulation mechanisms and *PRKD2* (12,13).

https://doi.org/10.59720/24-060

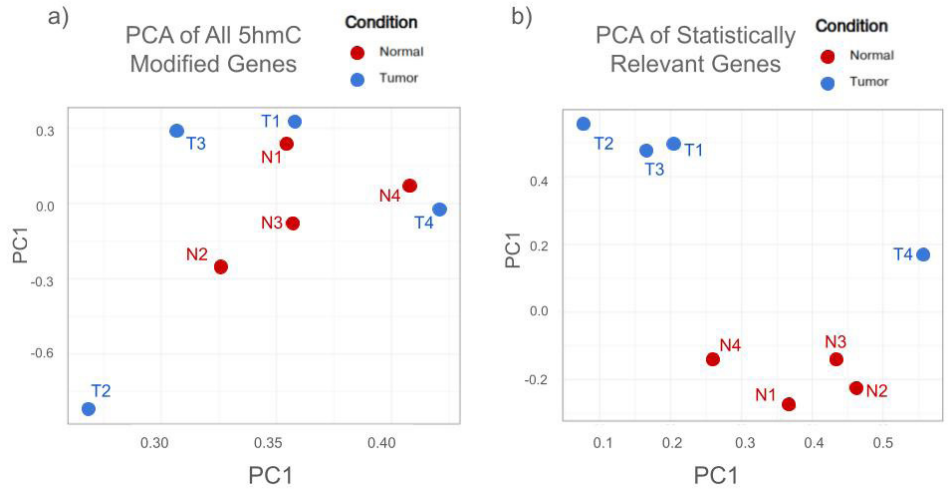

Figure 4: PCA plots to visualize the variance of 5hmC-containing genes. PCA was used to reduce the dimensionality of the dataset and highlight differences between tumor and normal samples based on 5hmC modification levels. The input data were derived from CMS-seq libraries, where the levels of 5hmC in each gene were quantified. Statistically significant genes were identified using a two-fold change threshold and a p-value < 0.05, adjusted using a Benjamini-Hochberg procedure for false discovery rate. Points labeled with the starting character "N" are plotted from normal samples and "T" from tumor samples. Points labeled with the same ending number are from samples from the same individual. (a) PCA plot using all 5hmC-modified genes. (b) PCA plot restricted to only using the most statistically relevant 336 5hmC-modified genes.

To further investigate how 5hmC modification regulates gene expression, we loaded our sample files into the Integrative Genomics Viewer (IGV) to visualize each sample's reads mapped to genomic regions (28). In a 5hmC concentrated area of *PRKD2* in chromosome 19, we found a comparable amount of 5hmC between the tumor samples (T1-T4) and normal samples (N1-N4). Interestingly, most of the 5hmC in the tumor samples at this snapshot within the gene body were shifted slightly forward compared to the healthy samples which suggested the location of 5hmC also plays a role in regulating gene expression (**Figure 7**).

HADHA encodes a subunit of the mitochondrial trifunctional protein (MTP), which is essential for breaking long-chain fatty acids in the mitochondria (29). Changes in metabolism and mitochondrial mutations can also influence cancer cell survival and proliferation (30). When evaluating *HADHA* expression levels, we found no significant difference between normal and tumor samples. The log2fold change was negligible, and the median values produced from GEIPA 2 were nearly level (**Figure 6b**). This suggests that HADHA does not show a significant difference in our current study between the two conditions.

AIPL1 has been found to play an important role in the eye and the functioning of photoreceptors, which detect light and enable vision in the retina (31). We note that GEIPA 2 indicates a much smaller expression level of *AIPL1*, most commonly only slightly above zero TPM (**Figure 6c**). Additionally, there is no significant change in transcription levels. Therefore, we see little correlation between *AIPL1* expression levels and head-neck cancer.

DISCUSSION

In this study, we utilized FFPE DNA samples, an underutilized but common form of DNA preservation. Throughout the study, we performed rigorous quality control and assessed DNA concentration to ensure usable libraries for sequencing. While FFPE samples are known to yield damaged DNA, our results demonstrate that FFPE DNA samples can be used for 5hmC-containing gene profiling using our improved CMS-seq method. Other studies have also found FFPE samples contain a sufficient amount of

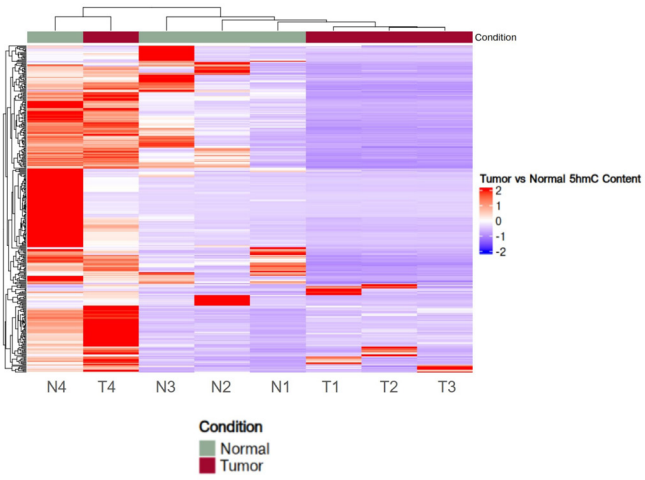

Figure 5: Comparative heatmap of 5hmC concentrations. The dendrograms (brackets) on the top and left sides of the heatmap represent hierarchical clustering of samples (columns) and genes (rows), respectively. Clustering determined based on the similarity of 5hmC levels between samples or genes, with closely related ones grouped together, to compare the 5hmC levels in various samples at the gene level. Each tissue sample is depicted by a column of color-coded strips that indicate the normalized concentration of 5hmC in genes using z-scores, represented by a row. The color gradient extends from blue (the lowest 5hmC concentration) to red (the highest 5hmC concentration), relative to the whole dataset.

https://doi.org/10.59720/24-060

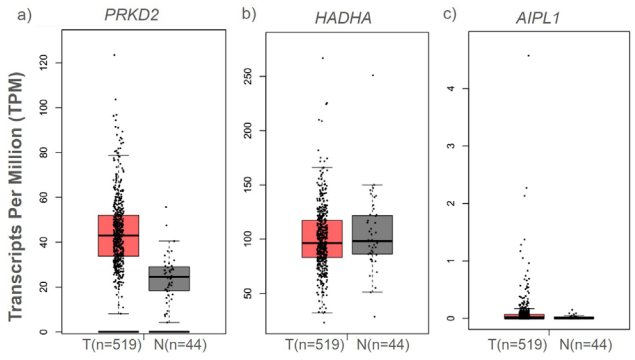

Figure 6: The gene expression levels of PRKD2, HADHA, and AIPL1 in HNC patients vs. healthy controls using publicly available databases. Gene expression levels in transcripts per million of (a) PRKD2, (b) HADHA, and (c) AIPL1 in 518 head-neck tumors (T, red box) compared to 44 paired normal samples (N, gray box) along x-axis. The comparison and the graph were automatically generated from the platform, Gene Expression Profiling Interactive Analysis (GEPIA 2, using the ANOVA test. The compared expressions of tumor and normal samples of the three genes were not shown to be statistically significant.

extractable DNA for NGS sequencing and that there is no significant quality difference from DNA extracted from FFPE samples stored for only 1-2 years or for more than a decade (16,17).

There is variety in 5hmC levels even within the normal and tumor tissue groups. These differences can result from a combination of many factors, including the patient's lifestyle, environment, habits, and other unique circumstances that are not limited to hereditary traits. While cancer is not the only factor that can change someone's epigenetic patterns, we can still identify a clear distinction in 5hmC levels between the healthy and tumor tissues of the same patients. One outlier out of four total samples may reflect the biological sample differences, and the probability of 5hmC as a biomarker to detect HNC, but not affect the overall robustness of the CMS-seq method. This study investigates the possibility of using CMS-seq to identify 5hmC as a potential biomarker for HNC detection. Our CMS-seq method can be extended to many FFPE samples to seek disease biomarkers for a wider dataset to increase accuracy and robustness in the future.

We identified three genes, *PRKD2, HADHA, and AIPL1*, as the most statistically significant, with 5hmC levels at least eight times smaller in tumor tissues than in their healthy counterparts. The identified genes based on 5hmC levels correlate to HNC, demonstrating the promising potential of 5hmC in distinguishing HNC's epigenetic landscape and in discovering new candidates for biomarker research. This would be useful for HNC treatment because of HNC's heterogeneous nature (5).

The low levels of 5hmC in these three genes may facilitate HNSCC development. The shifting in 5hmC enrichment location observed in *PRKD2* may help us understand the mechanisms and role of 5hmC in cancer biology in future research. At the same time, we examined the expression levels of the three identified genes in HNC patients, leveraging the publicly available databases TCGA and GTEx. Of the three genes we identified, *PRKD2* might play a role in tumor progression of HNC, while HADHA and AIPL1 do not appear to significantly influence tumor progression or suppression based on their expression levels. We found that the 5hmC level is not always positively correlated to the gene expression level, as in the case of PRKD2, underscoring the complexity of 5hmC regulatory function.

Nevertheless, we also note the limitations of this study. The small sample size limits the statistical significance of our work. We tested four individual's HNC tumor tissue and normal tissue samples in this pilot study, and larger-scale studies will be needed to draw more solid conclusions. Still, the lower level of 5hmC found in HNC tumor compared to surrounding healthy tissues suggests that 5hmC level itself may be a good biomarker for HNC. Furthermore, 5hmC can potentially

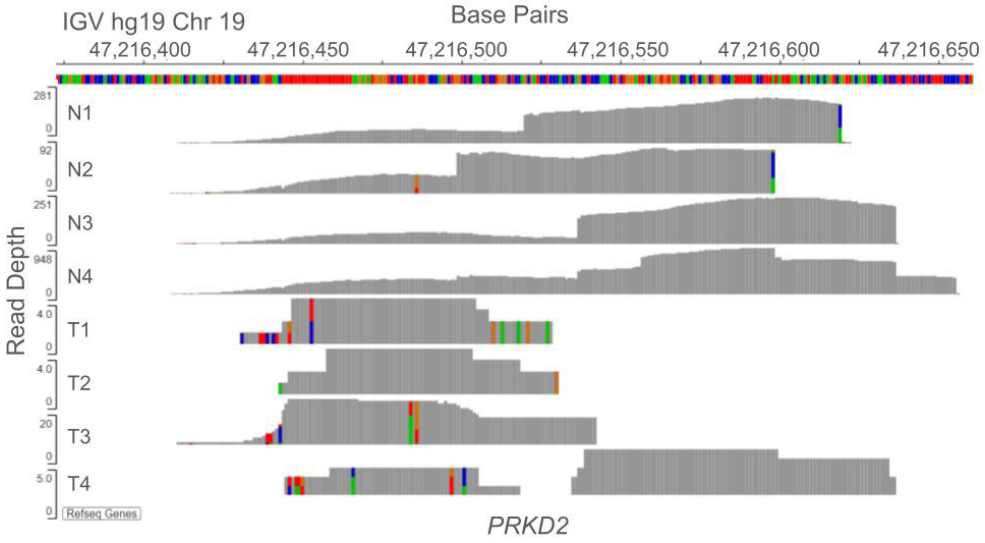

Figure 7: A snapshot of decreased 5hmC level in tumors within PRKD2. An example of a shifted pattern in PRKD2 5hmC-containing sequence using the human reference genome hg19 through Integrative Genomics Viewer (IGV). Samples listed from top to bottom with N = normal and T = tumor. Samples labeled with the same ending number came from the same individual. The bar heights reflect 5hmC coverage, with higher bars indicating more 5hmC. Colors represent nucleotide bases: red (T), blue (C), green (A), and orange (G).

https://doi.org/10.59720/24-060

identify biomarkers, but more work is needed to identify statistically robust biomarkers. Furthermore, our research was only focused on the 5hmC levels in the gene body so additional studies will be needed to identify biomarkers, such as a detailed examination of the gene's expression profiles. This project contributes to the continuously growing knowledge on 5hmC modification in cancer research and explores the traditionally unpopular use of FFPE samples in epigenetic sequencing. By addressing the limitations noted, future studies using 5hmC could advance our understanding of HNC's epigenetic landscape and improve treatment strategies by discovering novel biomarkers.

MATERIALS AND METHODS

Construction of 5hmC-enriched Next Generation Sequencing library

Pre-extracted genomic DNA was used from four pairs of human HNC tumors and adjacent normal tissue FFPE samples up to a decade old (unpublished, University of Chicago). From each, 50 ng of FFPE-isolated DNA was used as input. The DNA samples were fragmented and adapter-ligated using the KAPA Hyper Prep Kit (Roche, Indianapolis, IN). Then, the DNA was subjected to BS treatment at neutral pH conditions to avoid undesired C-to-U conversion, severe DNA damage, and ensure only 5hmC transformed into CMS. An anti-CMS serum was used to pull down 5hmC-containing DNA fragments from all samples through magnetic protein beads. The DNA fragments that did not contain CMS were discarded. The beads containing DNA identified with CMS were released by heating the samples at 95°C for five minutes. Since a 3'-adaptor was introduced, linear PCR was conducted to amplify the enriched DNA fragments to avoid the loss of DNA fragments with a very small number of copies and to maintain DNA quality. Then 5'-ligation and PCR was performed to amplify the fragment reads exponentially and complete libraries (**Figure 1**). The constructed libraries were then sent for NGS Illumina sequencing.

Quality control and processing

The eight libraries were paired-end sequenced to 100 bp using the Illumina NovaSeq 6000 sequencer and returned in the form of sixteen paired-end FASTQ files. All raw data processing and sequence alignment was done through a Linux command line server. The FASTQ files were run through FastQC to visualize the raw data quality. Using the FastQC reports, further data refining steps were determined to ensure quality during downstream analysis. The raw data files were quality-trimmed using Trim Galore, scripted to take paired-end files as input (32). Trim Galore defaults to cutting base pair reads using a base quality requirement of Phred score 20, representing a probability of 1 error in 100 reads and the autodetected adapter sequences. If sequences were determined to be too short after trimming for high-accuracy alignment, the entire sequence was discarded by Trim Galore. For every pair of input files, an output file containing 75% to 85% of the original data that passed the quality control filters was used for downstream analysis (**Table 1**). The reads were then aligned using Bowtie2 due to its ability to align reads with lengths greater than 50 bps (base pairs) to large reference genomes (33). The counts per gene were analyzed and differential 5hmC-modified genes were identified in tumor tissues compared to healthy tissues (**Figure 2**).

	T1	N1	T2	N2	T3	N3	T4	N4
# of Reads	578,273	842,161	608,263	1,898,870	596,359	1,497,731	6,772,565	8,050,082
Mapped ratio	93.45%	92.97%	92.19%	92.28%	92.96%	92.41%	97.16%	95.02%
Read length (mean)	75.7	72.45	74.38	75.01	76.79	73.54	75.22	73.64

Table 1: Overview of the eight NGS libraries after alignment using Qualimap. The counts of sequences acquired for each sample, alignment percentage to the human reference genome (hg19), and average sequence length. Points labeled with the starting character "N" are plotted from normal samples and "T" from tumor samples. Points labeled with the same ending number are from samples from the same individual.

Sequence alignment and read quantification

The processed sequences were aligned to the reference genome GRCh37 (hg19), obtained from the National Center for Biotechnology Information (NCBI), using Bowtie 2 (33). When aligning, one misalignment was allowed for flexibility without compromising accuracy. After alignment, the sequences were sorted using Samtools based on genome coordinates to prepare for read quantification (34). The number of reads that map to a gene was counted using the FeatureCounts tool, which uses existing annotations from the reference genome to assign reads and quantify 5hmC expression levels in gene bodies (35). The package DESeq2 was used for differential expression analysis to identify the genes showing significantly different 5hmC expression levels between each tumor and normal tissue sample pair (36). The Wald test was used by DESeq2 to calculate raw *p*-values (36).

Data analysis and visualization

To analyze the 5hmC profiling data, the CMS-seq libraries were processed by first identifying 14,606 genes with more than 50 total reads across all eight samples for downstream analysis. This step helped ensure statistical robustness and avoided bias toward genes with minimal 5hmC modifications. The z-scores were then calculated for the 5hmC levels of each gene, normalizing them to the dataset, which allowed 5hmC concentration levels across samples to be visualized in RStudio.

For dimensionality reduction and pattern recognition, Principal Component Analysis (PCA) was performed, using the 339 statistically significant genes with over a two-fold change in 5hmC levels between tumor and normal tissues. Tumor samples were differentiated from normal tissues using PCA by examining their clustering patterns along the principal components. Additionally, volcano plots were generated in RStudio to highlight genes with the most significant 5hmC modifications between tumor and normal tissues, and a *p*-adjusted value threshold of 0.05, calculated using the Benjamini-Hochberg procedure in DESeq2, was applied to improve statistical confidence (36).

Selected genes validation through public platforms

The expression levels of PRKD2, HADHA, and AIPL1 were investigated using two publicly available datasets.

https://doi.org/10.59720/24-060

One was The Cancer Genome Atlas (TCGA), which gives gene expression information for different cancer types. The other was The Genotype-Tissue Expression (GTEx), which gives gene expression information for normal tissues. Gene Expression Profiling Interactive Analysis (GEPIA 2) was used to assess differential gene expression across sample groups from the two datasets.

Received: February 15, 2024
Accepted: July 02, 2024
Published: March 11, 2025

REFERENCES

1. Barsouk, Adam, et al. "Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma." Medical Sciences, vol. 11, no. 2, June 2023, p. 42. https://doi.org/10.3390/medsci11020042.
2. Li, Qingfang, et al. "Targeted Therapy for Head and Neck Cancer: Signaling Pathways and Clinical Studies." Signal Transduction and Targeted Therapy, vol. 8, no. 1, 1, Jan. 2023, pp. 1–28. https://doi.org/10.1038/s41392-022-01297-0.
3. Gaździcka, Jadwiga, et al. "Epigenetic Modifications in Head and Neck Cancer." Biochemical Genetics, vol. 58, no. 2, Apr. 2020, pp. 213–44. https://doi.org/10.1007/s10528-019-09941-1.
4. Misawa, Kiyoshi, et al. "5-Hydroxymethylcytosine and Ten-Eleven Translocation Dioxygenases in Head and Neck Carcinoma." Journal of Cancer, vol. 10, no. 21, Aug. 2019, pp. 5306–14. https://doi.org/10.1007/s10528-019-09941-1.
5. Romanowska, Kamila, et al. "Head and Neck Squamous Cell Carcinoma: Epigenetic Landscape." Diagnostics, vol. 11, no. 1, Dec. 2020, p. 34. https://doi.org/10.3390/diagnostics11010034.
6. Breiling, Achim, and Frank Lyko. "Epigenetic Regulatory Functions of DNA Modifications: 5-Methylcytosine and Beyond." Epigenetics & Chromatin, vol. 8, no. 1, Dec. 2015, p. 24. https://doi.org/10.1186/s13072-015-0016-6.
7. Jang, Hyun Sik, et al. "CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function." Genes, vol. 8, no. 6, May 2017, p. 148. https://doi.org/10.3390/genes8060148.
8. Thomson, John P., and Richard R. Meehan. "The Application of Genome-Wide 5-Hydroxymethylcytosine Studies in Cancer Research." Epigenomics, vol. 9, no. 1, Jan. 2017, pp. 77–91. https://doi.org/10.2217/epi-2016-0122.
9. Shi, Dong-Qiao, et al. "New Insights into 5hmC DNA Modification: Generation, Distribution and Function." Frontiers in Genetics, vol. 8, July 2017, p. 100. https://doi.org/10.3389/fgene.2017.00100.
10. Cui, et al. "A Human Tissue Map of 5-Hydroxymethylcytosines Exhibits Tissue Specificity through Gene and Enhancer Modulation." Nature Communications. www.nature.com/articles/s41467-020-20001-w.
11. Li, Weiwei, and Min Liu. "Distribution of 5-Hydroxymethylcytosine in Different Human Tissues." Journal of Nucleic Acids, vol. 2011, June 2011, p. 870726. https://doi.org/10.4061/2011/870726.
12. Ecsedi, S., Rodríguez-Aguilera, J. R., & Hernandez-Vargas, H. (2018). 5-Hydroxymethylcytosine (5hmC), or How to Identify Your Favorite Cell. Epigenomes, 2(1), Article 1. https://doi.org/10.3390/epigenomes2010003.
13. He, Bo, et al. "Tissue-Specific 5-Hydroxymethylcytosine Landscape of the Human Genome." Nature Communications, vol. 12, no. 1, 1, July 2021, p. 4249. https://doi.org/10.1038/s41467-021-24425-w.
14. Hu, Xinlei, et al. "Integrated 5-Hydroxymethylcytosine and Fragmentation Signatures as Enhanced Biomarkers in Lung Cancer." Clinical Epigenetics, vol. 14, Jan. 2022, p. 15. https://doi.org/10.1186/s13148-022-01233-7.
15. Ding, Qian, et al. "Fecal Biomarkers: Non-Invasive Diagnosis of Colorectal Cancer." Frontiers in Oncology, vol. 12, Sept. 2022, p. 971930. https://doi.org/10.3389/fonc.2022.971930.
16. Nagahashi, Masayuki, et al. "Formalin-Fixed Paraffin-Embedded Sample Conditions for Deep next Generation Sequencing." The Journal of Surgical Research, vol. 220, Dec. 2017, pp. 125–32. https://doi.org/10.1016/j.jss.2017.06.077.
17. Kokkat, Theresa J., et al. "Archived Formalin-Fixed Paraffin-Embedded (FFPE) Blocks: A Valuable Underexploited Resource for Extraction of DNA, RNA, and Protein." Biopreservation and Biobanking, vol. 11, no. 2, Apr. 2013, pp. 101–06. https://doi.org/10.1089/bio.2012.0052.
18. Mathieson, William, and Geraldine A. Thomas. "Why Formalin-Fixed, Paraffin-Embedded Biospecimens Must Be Used in Genomic Medicine: An Evidence-Based Review and Conclusion." Journal of Histochemistry and Cytochemistry, vol. 68, no. 8, Aug. 2020, pp. 543–52. https://doi.org/10.1369/0022155420945050.
19. Srinivasan, Mythily, et al. "Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids." The American Journal of Pathology, vol. 161, no. 6, Dec. 2002, pp. 1961–71. https://doi.org/10.1016/S0002-9440(10)64472-0.
20. Li, Wenshuai, et al. "5-Hydroxymethylcytosine Signatures in Circulating Cell-Free DNA as Diagnostic Biomarkers for Human Cancers." Cell Research, vol. 27, no. 10, 10, Oct. 2017, pp. 1243–57. https://doi.org/10.1038/cr.2017.121.
21. Yu, M., Hon, G. C., Szulwach, K. E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J.-H., Jin, P., Ren, B., & He, C. (2012). Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. Cell, 149(6), 1368–1380. https://doi.org/10.1016/j.cell.2012.04.027.
22. Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W., & Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nature Protocols, 8(10), 1841–1851. https://doi.org/10.1038/nprot.2013.115.
23. Huang, Y., Pastor, W. A., Zepeda-Martínez, J. A., & Rao, A. (2012). The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. Nature Protocols, 7(10), 1897–1908. https://doi.org/10.1038/nprot.2012.103.
24. Pfeifer, Gerd P., Wenying Xiong, et al. "The Role of 5-Hydroxymethylcytosine in Human Cancer." Cell and Tissue Research, vol. 356, no. 3, June 2014, pp. 631–41. https://doi.org/10.1007/s00441-014-1896-7.
25. "Principal Component Analysis (PCA) Explained." Built In. www.builtin.com/data-science/step-step-explanation-

https://doi.org/10.59720/24-060

principal-component-analysis. Accessed 30 Dec. 2023.
26. Azoitei, Ninel, et al. "Protein Kinase D2: A Versatile Player in Cancer Biology." Oncogene, vol. 37, no. 10, 10, Mar. 2018, pp. 1263–78. https://doi.org/10.1038/s41388-017-0052-8.
27. Tang, Zefang, et al. "GEPIA2: An Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis." Nucleic Acids Research, vol. 47, no. W1, July 2019, pp. W556–60. https://doi.org/10.1093/nar/gkz430.
28. Robinson, James T., et al. "Integrative Genomics Viewer." Nature Biotechnology, vol. 29, no. 1, 1, Jan. 2011, pp. 24–26. https://doi.org/10.1038/nbt.1754.
29. Maeyashiki, Chiaki et al. "HADHA, the alpha subunit of the mitochondrial trifunctional protein, is involved in long-chain fatty acid-induced autophagy in intestinal epithelial cells." Biochemical and biophysical research communications vol. 484,3 (2017): 636-641. https://doi.org/10.1016/j.bbrc.2017.01.159.
30. Wallace, Douglas C. "Mitochondria and Cancer." Nature Reviews Cancer, vol. 12, no. 10, 10, Oct. 2012, pp. 685–98. https://doi.org/10.1038/nrc3365.
31. Sacristan-Reviriego, Almudena, and Jacqueline van der Spuy. "The Leber Congenital Amaurosis-Linked Protein AIPL1 and Its Critical Role in Photoreceptors." Advances in Experimental Medicine and Biology, vol. 1074, 2018, pp. 381–86. https://doi.org/10.1007/978-3-319-75402-4_47.
32. Krueger, Felix, et al. FelixKrueger/TrimGalore: V0.6.10 - Add Default Decompression Path. 0.6.10, Zenodo, 2 Feb. 2023. https://doi.org/10.5281/ZENODO.7598955.
33. Langmead, Ben, and Steven L. Salzberg. "Fast Gapped-Read Alignment with Bowtie 2." Nature Methods, vol. 9, no. 4, Mar. 2012, pp. 357–59. https://doi.org/10.1038/nmeth.1923.
34. Danecek, Petr, et al. "Twelve Years of SAMtools and BCFtools." GigaScience, vol. 10, no. 2, Jan. 2021, p. giab008. https://doi.org/10.1093/gigascience/giab008.
35. Liao, Yang, et al. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." Bioinformatics, vol. 30, no. 7, Apr. 2014, pp. 923–30. https://doi.org/10.1093/bioinformatics/btt656.
36. Love, Michael I., et al. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." Genome Biology, vol. 15, no. 12, Dec. 2014, p. 550. https://doi.org/10.1186/s13059-014-0550-8.

Copyright: © 2025 Li and Dai. All JEI articles are distributed under the attribution non-commercial, no derivative license (http://creativecommons.org/licenses/by-nc-nd/4.0/). This means that anyone is free to share, copy and distribute an unaltered article for non-commercial purposes provided the original author and source is credited.