Brush swab as a noninvasive surrogate for tissue biopsies in oral cancer patients to develop clinically translatable epigenetic biomarkers

Chi Tonglien Viet (cviet@llu.edu)
Loma Linda University  https://orcid.org/0000-0001-5078-7807

Xinyu Zhang
Yale University

Ke Xu
Yale University

Gary Yu
New York University

Kesava Asam
New York University

Carissa M. Thomas
University of Alabama at Birmingham

Nicholas F. Callahan
University of Illinois at Chicago

Coleen Doan
Loma Linda University

Paul C. Walker
Loma Linda University

Khanh Nguyen
Loma Linda University

Stephanie C. Kidd
Loma Linda University

Steve C. Lee
Loma Linda University

Anupama Grandhi
Loma Linda University

Clint T. Allen
NIDCD: National Institute on Deafness and Other Communication Disorders

Simon Young
University of Texas Health Science Center at Houston School of Dentistry: The University of Texas Health Science Center at Houston School of Dentistry

James C. Melville
University of Texas Health Science Center at Houston School of Dentistry: The University of Texas Health Science Center at Houston School of Dentistry

Jonathan W. Shum
University of Texas Health Science Center at Houston School of Dentistry: The University of Texas Health Science Center at Houston School of Dentistry

Dan T. Viet
Rocky Vista University

Alan S. Herford
Loma Linda University

Dylan F. Roden
Rutgers New Jersey Medical School

Manuel Lora Gonzalez
University of Alabama at Birmingham

Bradley E. Aouizerat
New York University

Research Article

Keywords: methylation biomarker, oral cancer, head and neck cancer, epigenetic biomarker, brush swab, brush biopsy, biomarker, MethylCap-Seq, methylation array

Posted Date: July 27th, 2021
Abstract

Background

Oral squamous cell carcinoma (OSCC) has poor survival rates. There is a pressing need to develop more precise risk assessment methods to tailor clinical treatment. Epigenome-wide association studies in OSCC have not produced a viable biomarker. These studies have relied on methylation array platforms, which are limited in their ability to profile the methylome. In this study, we use MethylCap-Seq (MC-Seq), a comprehensive methylation quantification technique, and brush swab samples, to develop a noninvasive, readily translatable approach to profile the methylome in OSCC patients.

Methods

Three OSCC patients underwent collection of cancer and contralateral normal tissue and brush swab biopsies, totaling 4 samples for each patient. Epigenome-wide DNA methylation quantification was performed using the SureSelectXT Methyl-Seq platform. DNA quality and methylation site resolution were compared between brush swab and tissue samples. Correlation and methylation value difference were determined for brush swabs vs. tissues for each respective patient and site (i.e., cancer or normal). Correlations were calculated between cancer and normal tissues and brush swab samples for each patient to determine the robustness of DNA methylation marks using brush swabs in clinical biomarker studies.

Results

There were no significant differences in DNA yield between tissue and brush swab samples. Mapping efficiency exceeded 90% across all samples, with no differences between tissue and brush swabs. The average number of CpG sites with at least 10x depth of coverage was 2,716,674 for brush swabs and 2,903,261 for tissues. Matched tissue and brush swabs had excellent correlation (r = 0.913 for cancer samples and r = 0.951 for normal samples). The methylation profile of the top 1,000 CpGs was significantly different between cancer and normal samples (mean p-value = 0.00021) but not different between tissues and brush swabs (mean p-value = 0.11).

Conclusions

Our results demonstrate that MC-Seq is an efficient platform for epigenome profiling in cancer biomarker studies, with broader methylome coverage than array-based platforms. Brush swab biopsy provides adequate DNA yield for MC-Seq, and taken together, our findings set the stage for development of a noninvasive methylome quantification technique for oral cancer with high translational potential.

Introduction

Each year 30,000 patients are diagnosed with oral cavity squamous cell carcinoma (OSCC), and unfortunately the incidence is on the rise [1–3]. Even for these early stage patients, the five-year survival rate is 60% [4]. Poor survival rates are in part due to inaccurate risk prediction. Early stage OSCC is primarily treated with surgical resection of the cancer, with or without adjuvant treatments such as an elective lymphadenectomy, radiation, or chemoradiation, for patients with high risk features. Currently, risk prediction to assign adjuvant treatment is entirely based on clinicopathologic information. Multiple retrospective and prospective studies have shown that these standard clinicopathologic factors have moderate accuracy with a concordance statistic (c-statistic) of 0.7 [4, 5]. The key to improving survival in OSCC lies in developing more accurate risk prediction methods, particularly in early stage patients. Although OSCC is a heavily epigenetically-regulated cancer [6], optimizing risk prediction using methylation features remains in its infancy. Methylation is one of the most frequent epigenetic changes in early oral carcinogenesis that is linked to cancer progression [6]. While several methylation studies in OSCC patients [6–16], including our own studies [7, 8], have highlighted differential methylation features between low and high risk patients, none of these studies have resulted in a clinically meaningful biomarker. Two main shortcomings of these previous studies are: 1) failure to use a clinically translatable array platform, and 2) failure to quantify methylation in real time, as cancer treatment is occurring.

With respect to the first challenge, the vast majority of methylation array studies in OSCC have used array-based platforms. While the Illumina Methylation 450K or EPIC array are the most commonly used platforms for epigenome-wide association studies (EWAS), CpG site quantification is restricted at an upper limit of 870,000 sites, and results from these platforms have not been converted into a clinically-accessible risk prediction tool. Furthermore, the EPIC array content is frequently updated to enrich for cancer-associated genes, making comparison across cohorts challenging. Methylation capture sequencing (MC-seq) has a scalable workflow that can quantify methylation in a small subset of genes or the entire genome using next generation sequencing (NGS), with a higher likelihood of clinical translation due to broader CpG coverage in a more agnostic manner while maintaining its resolution in samples with modest DNA quantities [17].

With respect to the second challenge, clinical translation of a biomarker requires measurement at the onset of treatment in order to determine risk and the need for treatment escalation. Waiting until after cancer removal for the formalin-fixed, paraffin-embedded (FFPE) tissues would limit clinical translatability. The oral cavity has the advantage of being readily accessible for sampling, not only with tissue biopsies, but also with noninvasive techniques. Herein, we determine methylation features using noninvasive brush swabs.

In this study, we hypothesize that brush swab biopsies serve as a robust noninvasive method to quantify cancer-specific methylation features. Using tissue and brush swab biopsies collected from OSCC patients at the time of surgery: 1) we determine the concordance between the methylation signature of cancer...
tissues and swabs vs. matched normal tissues and swabs using MC-Seq, and 2) we establish a workflow in which brush swabs and MC-seq are used at the time of diagnosis to establish a methylation signature that can be used to determine risk of mortality.

**Methods**

**Patient selection and data collection**

The patients were enrolled in a multi-institutional prospective clinical study in which biological samples and clinicopathologic information were collected. Collection of clinical data and samples was approved by the Institutional Review Board at each institution, which included Loma Linda University (LLU), University of Illinois Chicago (UIC), and University of Alabama at Birmingham (UAB). Patients were eligible if they were ≥ 18 years of age, had biopsy-proven squamous cell carcinoma of oral cavity sub-sites, including oral tongue, maxillary and mandibular gingiva, hard palate, floor of mouth, buccal mucosa, and lip mucosa, and no previous treatment of OSCC. Clinical and pathologic stages were recorded based on the American Joint Committee on Cancer (AJCC) Eighth Edition Staging Manual [18]. We collected the following information from the chart review: age, sex, race, smoking and alcohol use, staging, tumor location, pathologic characteristics, and treatment modalities received in addition to tumor ablation. Biological samples collected at the time of surgery include flash-frozen cancer and contralateral normal tissue, and brush swab biopsies of the cancer and contralateral normal site. Samples were stored in -80°C. A total of 3 patients were randomly chosen from the ongoing prospective clinical study for the current study.

**Nucleic acid extraction and sample preparation**

DNA was extracted from the fresh-frozen tissue and brush swabs of the cancer and contralateral normal side of 3 patients, totaling 12 samples (4 samples per patient). Genomic DNA quality was determined by spectrophotometry and concentration was determined by fluorometry. DNA integrity and fragment size were determined using a microfluidic chip run on an Agilent Bioanalyzer.

**MC-seq target enrichment library prep**

Indexed paired-end whole-genome sequencing libraries were prepared using the SureSelect XT Methyl-Seq kit (Agilent). Genomic DNA was sheared to a fragment length of 150–200 bp using the Covaris E220 system. Fragmented sample size distribution was determined using the Caliper LabChip GX system (PerkinElmer). Fragmented DNA ends were repaired with T4 DNA Polymerase and Polynucleotide Kinase and “A” base was added using Klenow fragment followed by AMPure XP bead-based purification (Beckman Coulter). The methylated adapters were ligated using T4 DNA ligase followed by bead purification with AMPure XP. Quality and quantity of adapter-ligated DNA were assessed with the Caliper LabChip GX system. Samples were enriched for targeted methylation sites by using the custom SureSelect Methyl-Seq Capture Library. Hybridization was performed at 65°C for 16 h using a thermal cycler. Once the enrichment was completed, the samples were mixed with streptavidin-coated beads (Thermo Fisher Scientific) and washed with a series of buffers to remove non-specific DNA fragments. DNA fragments were eluted from beads with 0.1 M NaOH. Unmethylated C residues of enriched DNA underwent bisulfite conversion using the EZ DNA Methylation-Gold Kit (Zymo Research). The SureSelect enriched and bisulfite-converted libraries underwent PCR amplification using custom made primers (IDT). Dual-indexed libraries were quantified by quantitative polymerase chain reaction (qPCR) with the Library Quantification Kit (KAPA Biosystems) and inserts size distribution was assessed using the Caliper LabChip GX system.

**Flow cell preparation and sequencing**

Samples were sequenced using 100 bp paired-end sequencing on an Illumina HiSeq NovaSeq according to Illumina protocol. A positive control (prepared bacteriophage Phi X library) was added into every lane at a concentration of 0.3% to assess sequencing quality in real time.

**Preprocessing and quality control**

Signal intensities were converted to individual base calls during each run using the system’s Real Time Analysis software. Sample de-multiplexing was performed using Illumina's CASAVA 1.8.2 software suite. The sample error rate was required to be less than 1% and the distribution of reads per sample in a lane to be within reasonable tolerance. Signal data quality were examined using FastQC (ver. 0.11.8). Adapter sequences and fragments with poor quality were removed by Trim_galore (ver. 0.6.3_dev). Bismark pipelines (ver. v0.22.1_dev) were used to align the reads to the bisulfite human genome (hg19) with default parameters.[19] Sample alignment to the human genome was performed using bowtie 2 (ver. 2.3.5.1). Quality-trimmed paired-end reads were converted into a bisulfite forward (C→T conversion) or reverse (G→A conversion) strand read. Duplicated reads were removed from the Bismark mapping output and CpG extracted. All CpG sites were grouped by sequencing coverage (i.e., read depth); CpG sites with coverage ≥ 10x depth were retained for analysis to ensure high MC-Seq data quality. Genes were annotated using Homer annotatePeaks.pl.

**Comparison of methylation between tissue and brush swab biopsies**

Pearson correlations were calculated between tissue and brush biopsy samples of matched anatomic sites. Pearson correlation and absolute difference were calculated among common CpG sites between the matched tissue and brush biopsies. Scatterplots were rendered showing the correlation of β values from all CpG sites measured by MC-eQ. Separate scatterplots were rendered showing the concordance of these CpG sites between tissues and brush swabs for the cancer sites and the normal sites. Student t-tests were performed to compare β values between cancer and normal groups or tissue and brush swab groups. The most significant 1,000 CpGs features in cancer vs. normal groups were selected. Based on these results the -log10(t-test p-value) was calculated for each of the 1,000 CpG sites to compare the degree of divergence in the significance of the test statistics for these 1,000 CpG between 1) cancer vs. normal and 2) tissue vs. brush swabs.

**Statistical analyses**

Statistical analyses were performed in R environment (v. 4.1.0).
Results

Patient cohort characteristics and DNA quality

Clinicopathologic information for the 3 enrolled patients are detailed in Table 1. The 3 patients comprised both early and late stage OSCC (stage I and IV), as well as varying tobacco and alcohol consumption habits. Patients were 49 and 68 years old. Two patients were male and one was female. All patients were white, non-Hispanic. Cancer and contralateral normal tissue and brush swab biopsies collected at the time of surgery underwent DNA extraction, with the yield and quality shown in Table 2. With a total input volume of 30µL for each sample, total input for tissue DNA ranged from 187ng to 660ng, and an average of 390ng. Total input for swab DNA ranged from 51ng to 1998ng, with an average of 532ng. The input range was consistent with our previous study demonstrating reproducible CpG site quantification using MC-Seq across this range [17].

Table 1
Patient demographic characteristics.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Sex</th>
<th>Race</th>
<th>Tobacco use, pack years</th>
<th>Alcohol, drinks/wk</th>
<th>Site</th>
<th>TNM</th>
<th>Stage</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>F</td>
<td>White</td>
<td>Never</td>
<td>Never</td>
<td>Tongue</td>
<td>T1N0M0</td>
<td>I</td>
<td>Moderate</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>M</td>
<td>White</td>
<td>Former, 53</td>
<td>Former, 24</td>
<td>Tongue</td>
<td>T4aN0M0</td>
<td>IV</td>
<td>Moderate</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>M</td>
<td>White</td>
<td>Current, 72</td>
<td>Current, 14</td>
<td>Mandible</td>
<td>T4bN3bM0</td>
<td>IV</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Table 2
Characteristics of genomic DNA used as input for sequencing of tissue and brush swab biopsies.

<table>
<thead>
<tr>
<th>Sample</th>
<th>DNA concentration ng/µl</th>
<th>A260</th>
<th>A280</th>
<th>260/280</th>
<th>260/230</th>
<th>gDNA input, ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>1C swab</td>
<td>3.96</td>
<td>0.01</td>
<td>-0.004</td>
<td>-2.99</td>
<td>-0.18</td>
<td>118.8</td>
</tr>
<tr>
<td>1C tissue</td>
<td>22.00</td>
<td>4.45</td>
<td>2.210</td>
<td>2.01</td>
<td>2.18</td>
<td>660.0</td>
</tr>
<tr>
<td>1N swab</td>
<td>1.70</td>
<td>-0.05</td>
<td>-0.048</td>
<td>0.95</td>
<td>0.33</td>
<td>51.0</td>
</tr>
<tr>
<td>1N tissue</td>
<td>6.24</td>
<td>0.44</td>
<td>0.220</td>
<td>2.01</td>
<td>2.75</td>
<td>187.2</td>
</tr>
<tr>
<td>2C swab</td>
<td>4.24</td>
<td>0.01</td>
<td>-0.027</td>
<td>-0.27</td>
<td>-0.09</td>
<td>127.2</td>
</tr>
<tr>
<td>2C tissue</td>
<td>11.40</td>
<td>1.00</td>
<td>0.463</td>
<td>2.16</td>
<td>2.85</td>
<td>342.0</td>
</tr>
<tr>
<td>2N swab</td>
<td>6.32</td>
<td>0.07</td>
<td>0.014</td>
<td>5.14</td>
<td>-1.45</td>
<td>189.6</td>
</tr>
<tr>
<td>2N tissue</td>
<td>8.00</td>
<td>0.62</td>
<td>0.290</td>
<td>2.12</td>
<td>3.04</td>
<td>240.0</td>
</tr>
<tr>
<td>3C swab</td>
<td>66.60</td>
<td>1.82</td>
<td>0.971</td>
<td>1.87</td>
<td>2.86</td>
<td>1998.0</td>
</tr>
<tr>
<td>3C tissue</td>
<td>21.80</td>
<td>2.16</td>
<td>1.314</td>
<td>1.65</td>
<td>0.80</td>
<td>654.0</td>
</tr>
<tr>
<td>3N swab</td>
<td>23.60</td>
<td>0.45</td>
<td>0.217</td>
<td>2.06</td>
<td>4.70</td>
<td>708.0</td>
</tr>
<tr>
<td>3N tissue</td>
<td>8.48</td>
<td>0.39</td>
<td>0.191</td>
<td>2.05</td>
<td>5.66</td>
<td>254.4</td>
</tr>
</tbody>
</table>

MC-Seq mapping efficiency assessment

Table 3 details the mapping efficiency for each biological sample. Using MC-Seq sequences mapped to the reference genome with an average mapping efficiency of 90% across all samples. There were no significant differences in mapping efficiency between tissues and brush swab samples (Fig. 1C). The average difference in mapping efficiency between the paired brush swabs and tissues was minimal, at -0.567%, in favor of tissue samples, with a range of -1.9 to 1.7%. The majority of methylated C’s appeared in a CpG context. We graphed the depth of read for each CpG across all queried CpGs and demonstrated an inflection point at 10x coverage (Fig. 1A). This finding was similar to our previous technical validation study, in which the majority of CpG sites exhibited at least 10x coverage [17]. We therefore applied this cutoff, focusing our analysis on CpG sites with at least 10x coverage. Average number of CpGs with at least 10x coverage was 2,716,674 for swab samples and 2,904,261 for tissue samples, with no significant difference between the two sample types, which is in excess of 3-fold greater CpGs interrogated than the most commonly used tool to measure the DNA methylome, the Illumina EPIC array. Figure 1B indicates the number of CpGs with at least 10x coverage for each of the 12 individual samples.
Distribution of methylome regions

We determined the distribution of CpG sites profiled by MC-Seq among the CpG sites successfully measured at 10X depth of read or greater overlapping across all 12 samples (3,566,843 CpGs). Figure 1D demonstrates that 36% were in introns, 26% were in promoters, 19% were in exons, and 19% were in intergenic regions. Overall, MC-Seq provided more robust coverage of functional gene regions in the methylome than typically provided by the EPIC array, detecting ten-fold more CpG sites in promoter regions and exons than the EPIC array. We determined that 484,697 CpGs from the EPIC array, the majority of which were also found on the 450K (396,409 CpG) were probed by MC-Seq with at least 10x coverage. While the breakdown of these CpGs was 33% intron, 15% exon, and 19% intergenic, the total number of CpGs in the functional gene regions was proportionally lower owing to the more limited context C’s in CpG methylated context C’s in CHG context C’s in CHH context CpG > 1%  

<table>
<thead>
<tr>
<th>Name</th>
<th>Mapping efficiency (%)</th>
<th>Difference between swab and tissue (%)</th>
<th>Sequence pairs analyzed in total</th>
<th>Number of paired-end alignments with a unique best hit</th>
<th>Duplicate (%)</th>
<th>Sequence pairs after removing duplicate</th>
<th>Total number of C’s in CpG context</th>
<th>Total methylated C’s in CHG context</th>
<th>Total methylated C’s in CHH context</th>
<th>CpG dep cov &gt; 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1C swab</td>
<td>90.7</td>
<td>0.1</td>
<td>42303352</td>
<td>38369629</td>
<td>52.22</td>
<td>18333433</td>
<td>749965581</td>
<td>37720278</td>
<td>2408697</td>
<td>5995678</td>
</tr>
<tr>
<td>1C tissue</td>
<td>90.6</td>
<td></td>
<td>34732207</td>
<td>31480976</td>
<td>28.42</td>
<td>22534549</td>
<td>927088792</td>
<td>44803115</td>
<td>2908059</td>
<td>7258418</td>
</tr>
<tr>
<td>1N swab</td>
<td>89.3</td>
<td>-0.7</td>
<td>35997131</td>
<td>32160165</td>
<td>69.49</td>
<td>9811824</td>
<td>392619063</td>
<td>19593941</td>
<td>1231170</td>
<td>3078167</td>
</tr>
<tr>
<td>1N tissue</td>
<td>90.4</td>
<td></td>
<td>38013580</td>
<td>34357686</td>
<td>50.99</td>
<td>16840023</td>
<td>701455939</td>
<td>36205354</td>
<td>2320490</td>
<td>5789438</td>
</tr>
<tr>
<td>2C swab</td>
<td>91.0</td>
<td>1.7</td>
<td>49206194</td>
<td>44795293</td>
<td>54.25</td>
<td>20495859</td>
<td>825691357</td>
<td>41320643</td>
<td>2723999</td>
<td>6792423</td>
</tr>
<tr>
<td>2C tissue</td>
<td>89.3</td>
<td></td>
<td>48990997</td>
<td>43728505</td>
<td>39.71</td>
<td>26364632</td>
<td>1095985402</td>
<td>53089834</td>
<td>3494884</td>
<td>8728001</td>
</tr>
<tr>
<td>2N swab</td>
<td>86.7</td>
<td>-1.9</td>
<td>40037338</td>
<td>34712643</td>
<td>48.22</td>
<td>17975067</td>
<td>745570397</td>
<td>35629449</td>
<td>2300915</td>
<td>5877073</td>
</tr>
<tr>
<td>2N tissue</td>
<td>88.6</td>
<td></td>
<td>37524719</td>
<td>33261053</td>
<td>48.50</td>
<td>17128931</td>
<td>713630671</td>
<td>35589419</td>
<td>2277861</td>
<td>5673182</td>
</tr>
<tr>
<td>3C swab</td>
<td>90.7</td>
<td>-0.8</td>
<td>36703056</td>
<td>33280675</td>
<td>26.89</td>
<td>24331566</td>
<td>1000848080</td>
<td>50517501</td>
<td>3160407</td>
<td>7841535</td>
</tr>
<tr>
<td>3C tissue</td>
<td>91.5</td>
<td></td>
<td>39728186</td>
<td>36357410</td>
<td>28.90</td>
<td>25848885</td>
<td>1069409107</td>
<td>52603435</td>
<td>3264928</td>
<td>8042519</td>
</tr>
<tr>
<td>3N swab</td>
<td>88.3</td>
<td>-1.8</td>
<td>52334030</td>
<td>46226126</td>
<td>34.76</td>
<td>30156313</td>
<td>1221133676</td>
<td>62639138</td>
<td>3840377</td>
<td>9388785</td>
</tr>
<tr>
<td>3N tissue</td>
<td>90.1</td>
<td></td>
<td>41553671</td>
<td>37445490</td>
<td>39.68</td>
<td>22588890</td>
<td>950370928</td>
<td>50055981</td>
<td>3072187</td>
<td>7486332</td>
</tr>
<tr>
<td>Average swab</td>
<td>89.45</td>
<td>-0.567</td>
<td>42763517</td>
<td>38257422</td>
<td>48.00</td>
<td>20184010</td>
<td>822638026</td>
<td>41236825</td>
<td>2610928</td>
<td>6495610</td>
</tr>
<tr>
<td>Average tissue</td>
<td>90.00</td>
<td></td>
<td>40090560</td>
<td>36105187</td>
<td>39.00</td>
<td>21884318</td>
<td>909656807</td>
<td>45391190</td>
<td>2889735</td>
<td>7162982</td>
</tr>
</tbody>
</table>

Correlation between brush swab and tissue biopsies from matched anatomic sites

Overall, the correlation among CpG site methylation across all samples was high, all exceeding 90%. The average correlation between tissue and brush swabs (n = 12) among all CpG sites shared among the entire sample (cancer + control) (s = 3,566,843) was 93.2% (95% confidence interval: 93.23%, 93.25%). The average correlation between tissue and brush swabs (n = 6) among all CpG sites shared among cancer samples was 91.3% (95% confidence interval: 91.32%, 91.35%). The average correlation between tissue and brush swabs (n = 6) among all CpG sites shared among normal samples was 95.1% (95% confidence interval: 95.13%, 95.14%). A scatterplot of the CpGs with 10x coverage was generated for the cancer samples and the normal samples separately, demonstrating high concordance between tissue and brush swabs (Fig. 2A and 2B).

The top methylation features are differentially methylated between cancer and normal samples, but not between tissues and brush swabs

We focused on the top 1,000 most variable methylation features between cancer and normal samples, which would be expected to differ considerably less between tissue and brush swab sampling methods. The p-values for each test of difference in CpG methylation by t-test were expressed as -log_{10}(p-value), and averaged 3.67 (i.e., p = 0.00021) between cancer vs. normal. The same CpG sites were not differentially methylated, with an average -log_{10}(p-value) = 0.96 (i.e., p = 0.11) between tissue vs. brush swabs (Fig. 2C). The results suggest that brush swabs are a clinically viable surrogate for tissue biopsies.
**Discussion**

MC-seq is a scalable, CLIA-approvable methylation assay that is currently not widely used in cancer research

EWAS studies in cancer patients have identified interindividual variability in the epigenome, and the recent availability of affordable EWAS technologies have led to a rapid increase in epigenetic biomarker studies aimed at identifying differential methylation features that could be predictive of clinical outcome. The most commonly used platforms are array-based, like the Illumina Human 450K and Infinium MethylationEPIC arrays, which provide limited coverage of CpG sites across the epigenome. Whole genome bisulfite sequencing (WGBS) is the most comprehensive method for genome profiling, capturing 28 million CpGs. However, the cost, intensive workflow, and need for high quality and quantity of DNA input significantly limit its clinical translatability, particularly in cancer treatment. MC-Seq has emerged as a promising intermediary between array-based platforms and WGBS, using NGS to capture significantly more CpGs than array-based platforms, while having the advantage of being more high-throughput and affordable than WGBS. We and others have compared CpG coverage and efficiency of different methylation quantification platforms [17, 20, 21]. A recent publication from our group has demonstrated that MC-Seq is a more reliable and efficient platform for epigenome profiling than array-based platforms like the EPIC array. When the EPIC array and MC-Seq were compared in peripheral blood mononuclear cell samples, MC-Seq captured significantly more CpGs in coding regions and CpG islands than the EPIC array. The EPIC array captured 846,464 CpG sites per sample, whereas MC-Seq captured 3,708,550 CpG sites per sample. Of the 472,540 CpG sites captured by both platforms, there was high correlation \( r = 0.98-0.99 \) in methylation status [17]. Moreover, while the EPIC array is enriched for genes with known roles in carcinogenesis, MC-Seq quantifies methylation in a more agnostic manner and profiles~4 times more CpGs than the EPIC array, allowing for a higher chance of discovering novel epigenetic modifications in cancer. Furthermore, the coverage areas within each gene were more comprehensive than the EPIC array and other commonly used methylation analysis techniques, like PCR or pyrosequencing. Herein, we demonstrated that MC-Seq captured significantly more CpG sites within functional gene regions, owing to the higher overall profiling capability of this technique.

**Oral SCC is an epigenetically-regulated cancer with promising methylation biomarker candidates**

Methylation studies on OSCC patients [6–16] including our own studies [7, 8] have demonstrated that methylation is a common event and highlighted specific genes for mechanistic studies. For example, a EWAS using the Illumina Human 450K array on 108 head and neck SCC patients of multiple sub-sites including oral cavity identified hypermethylation and inactivation of key tumor suppressor genes [9]. Clinical translation of these methylation biomarker studies has been limited due to: 1) combining OSCC with other head and neck cancer sub-sites (i.e., oropharynx, hypopharynx, larynx), which creates a heterogeneous cohort that fails to recognize OSCC as a distinct clinical disease, and 2) relying solely on array-based platforms, which query a limited number of CpGs. As a result, none of these studies have produced a methylation biomarker with high prognostic performance.

In addition to being a distinct clinical subsite from other head and neck sites, the oral cavity is an easily accessible anatomic site for non-invasive biopsy techniques. Clinical translation of a biomarker requires that it can be measured during treatment. Waiting until after tumor removal for the formalin-fixed, paraffin-embedded (FFPE) tissues delays potentially necessary treatment. Researchers have used both saliva and brush swabs to noninvasively sample OSCC cells at the time of diagnosis. In our own studies, we have used saliva to identify methylation biomarkers of OSCC. We demonstrated that a multi-gene panel could be constructed using either a methylation array or Methylight, a polymerase chain reaction (PCR) technique [7, 8]. However, we and others have shown that concordance of methylation between saliva and cancer tissue is highly variable [22, 23].

**Brush swabs and MC-Seq represent a noninvasive method to quantify methylation biomarkers**

Our approach of using brush swabs and MC-Seq to determine the methylation signature at the time of diagnosis has a high potential for clinical translatability. We demonstrated in this study that brush swab and tissue biopsies from matched sites had highly correlated methylation signatures. Furthermore, the DNA quality and quantity from brush swab samples were adequate to perform MC-Seq. Mapping efficiency was equivalent between tissues and brush swabs. Given the high correlation between the paired tissues and brush swabs, and the satisfactory DNA yield, brush swabs could serve as a clinically robust surrogate to tissue biopsies. One previous study has assessed the reliability of brush swab DNA for MC-Seq compared to the Human 450K array [20], drawing similar conclusions to our study [17] that MC-Seq offered broader coverage of CpG sites and that sample-based correlation was high \( r = 0.98 \) between the two platforms. However, they did not compare brush swab to underlying tissue collection. To our knowledge our study represents the first to directly compare the epigenome-wide signature of matched brush swabs and tissues, with the results having important implications in OSCC biomarker research.

**Conclusions**

Our study establishes a workflow for a large-scale clinical study using brush swab samples and MC-Seq to noninvasively determine the methylation signature of OSCC patients at the time of diagnosis, which could be used to establish risk stratification schemes.

**Abbreviations**

EWAS: epigenome-wide association study; MC-Seq: MethylCap-Seq; OSCC: oral squamous cell carcinoma

**Declarations**
References


Figures

(A) We compared depth of coverage in all CpGs and determined an inflection point at 10x coverage. (B) Using 10x read depth as a cutoff, we determined the number of quantified CpG sites in each sample. Average number of quantified CpGs meeting our criteria was 2,716,674 for swab samples and 2,904,261 for tissue samples, with no significant difference between the two sample types. (C) The average mapping efficiency was 89.45% for brush swabs and 90% for tissues, with no significant difference between the two sampling methods. (D) The pie charts detail the relative genic locations of the CpGs profiled by MC-Seq (left) and CpGs covered by the EPIC array that were profiled (right). MC-Seq provided more robust coverage of functional gene regions than the EPIC array.
The scatterplots demonstrate the correlation between tissue and brush swab biopsies for (A) cancer and (B) normal sites of the 3 patients. The correlation values are noted. (C) By focusing on the top 1,000 most variable methylation features between cancer and normal samples quantified with MC-Seq, we determined the methylation difference between different sample types, visualized using box plots (median, quartiles, maximum and minimum whiskers). The p-values for each test of difference in CpG methylation by t-test were expressed as -log10(p-value), which had a median of 3.67 (i.e., p=0.00021) between cancer vs. normal. The same CpG sites were not differentially methylated [average -log10(p-value) = 0.96 (i.e., p=0.11)] between tissue vs. brush swabs, suggesting that brush swabs are a viable surrogate of tissue biopsy.