**Read mapping**

Low-pass WGBS data will be processed as our previous publication[1]. We first trimmed all paired-end (PE) or single-end (SE) fastq files using trim-galore version 0.6.5 to remove low-quality bases and biased read positions. Next, the reads were encoded to map to a three-letter genome via conversion of all cytosines to thymidines or all guanines to adenines for the reads that seemed to be from the reverse-complement strand. Then the reads were mapped using Bowtie 2 (version [2.3.5.1](https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.5.1)) and Bismark (version 0.22.3), with the default setting to both the Watson-strand- and Crick-strand-converted genomes. The alignments with mapping-quality scores of < five were discarded, and only reads with a higher best-mapping-quality score in either the Watson or Crick strand were kept. Finally, the encoded read sequences were replaced by the original read sequences in the final BAM files. Overlapping paired-end reads were also clipped with the bamUtils clipOverlap function.

**Long-region mDELFI**

We applied similar approach to calculate mDELFI as the Dr. Cristiano’s previous publication[2] and replace WGS data with WGBS (bisulfite sequencing) data. To account for biases in coverage attributable to GC content of the genome, we applied locally weighted scatterplot smoothing (LOWESS, also known as LOESS) regression analysis with a span setting of 0.75 to the scatterplot of average fragment GC versus coverage calculated for each 100-kb bin. This LOESS regression was performed separately for short and long fragments to account for possible differences in GC effects on coverage in plasma by fragment length—an approach loosely motivated by a previous study. We subtracted the predictions for short and long coverage explained by GC from the LOESS model, obtaining residuals for short and long that were uncorrelated with GC. We returned the residuals to the original scale by adding back the genome-wide median short and long estimates of coverage. This procedure was repeated for each sample to account for possible differences in GC effects on coverage between samples. To reduce the feature space and noise further, we calculated the total GC-adjusted coverage in 5-Mb bins. To compare the variability of fragment lengths from healthy subjects to fragments in patients with cancer, we calculated the standard deviation of the short to long fragmentation profiles for each individual. We compared the median of the standard deviations in the two groups by a Wilcoxon rank-sum test.

**Prediction analysis between cancer and normal based on low-pass mDELFI data**

Validated ML analytical approaches will be applied in identifying early detection of dysplasia and malignant transformation associated with oral cavity and oro-pharyngeal cancers. Potential informatics approaches to conduct analysis includes key supervised ML approaches including Naïve Bayes, Logistic Regression, Decision Trees, Multi-layer perceptron-Artificial Neural Network and others to see which algorithms perform optimally or categorization applying support vector machines. ML algorithm will be developed and executed by using machine learning libraries (R version 3.6.2). ML models will be trained, tested and evaluated using 10 fold cross validation. Predicting will be treated as a classification problem, where the datasets will be sorted into two categories based on patient diagnosis of oral cavity cancer or oropharyngeal cancer. Performance of various ML algorithms will be evaluated and compared using the following metrics: total accuracy, recall, sensitivity and specificity. To select the best performing model Receiver Operating Characteristics (ROC) curve, area under ROC (AUC) and Recall-Precision curves for each ML algorithm will be plotted.

We initiated **mDELFI** method developing and deposited into github <https://github.com/Shicheng-Guo/mDELLFI> and will be continuously updated when any progression were made.

1. Guo, S., et al., *Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.* Nat Genet, 2017. **49**(4): p. 635-642.

2. Cristiano, S., et al., *Genome-wide cell-free DNA fragmentation in patients with cancer.* Nature, 2019. **570**(7761): p. 385-389.