**Figures Legends**

**Figure 1.** **Distribution of cellular DNA contents in exfoliative cytology experience “*please ask Yao to proof read A - C*”** **(A)** Selected cells with abnormally high DNA contents above D.I. value > 2.3. **(B)** Scatter plots of cells in the in exfoliative cytology study, y-axis indicates the area of captured nucleus image and x-axis indicates the corresponding DNA Index (D.I.) values. **(C)** Distribution histogram of D.I. values of all captured nucleus image. **(D)** Distribution histogram of D.I. values of simulated three cell populations and mixture of three. Red density was simulated from normal distribution (µ=1.001, σ=0.19); green density was simulated from normal distribution (µ=2.002, σ=0.25); blue density was simulated from normal distribution (µ=2.300, σ=0.5); black density was the mixture of three populations at ratio: 0.893:0.092:0.05.

**Figure 2. Expert-guided data transformation and reconstruction (EdTAR) work flow:** The EdTAR data process starts with DNA Index (D.I.) values. Briefly, the density of all D.I. values was explored and all candidate peaks were identified from the smoothed curve. With the expert-guided major parameters, i.e. theoretical mean of diploid or tetraploid cells, clinical OSCC diagnosis D.I. threshold, etc., diploid cell population was extracted and further filtered (if more than one populations were detected). Then, same procedure was applied onto existing tetraploid cell population leaving all the remaining cells of the aneuploidy population. Using the same expert-guided parameters (i.e. missing ratios of candidate populations) reconstruct new variables (of D.I.) across a wide rage [0 – 8] using the discrete density at each interval. The newly constructed data was used in training the statistical model and establish an Oral Cancer Risk Index (OCRI) panel.

**Figure 3. EdTAR processing results on three candidate samples clinically differentiated classes: Normal, OLK, and OSCC.**  (A-C) was from a normal sample, (D-F) was from an OLK sample, and (G-I) was from an OSCC sample. All density plots have x-axis of DNA Index (D.I.) value, where y-axis as density. Vertically, 3A, 3D and 3G showed density plot of all available D.I. values from each sample respectively. In the normal sample (3A), a major peek located at 0.995 represents the mean of diploid cell population, where another small peaks (D.I. = 0.594) was a minor hump possibly due to measurement error from the image process. In the OLK sample (3D), a major peek located at 0.798, represents the mean of diploid cell population. It deviated toward the left from “1” owing to the smoothing algorithm and it left another minor peak at 1.25. Three more almost invisible peaks were located at 1.75, 2.22, and 2.74 respectively. It was further indicated that the first cell population consist the main density (3590 cells). In the OSCC sample (3G), a major peek located at 1.02, represents the mean of diploid cell population. Another obvious peak was located at 1.79, which was deemed to represent the tetraploid population. The second peak represented the mean of the tetraploid population, which again deviated toward the left from “2” owing to the smoothing algorithm. Three more almost invisible peaks were located at 3.25, 3.57, and 3.99 respectively. These were deems to represent the aneuploidy cell population. The second column three plots showed the net results from EdTAR process, which was to reveal the hidden signals. (3B) was same as (3A) since only a single diploid population was identified and cleaning was unnecessary. (3E) and (3H) clearly showed the benefit from the EdTAR process where the tetraploid cell populations were showed and all possible candidate aneuploidy cells (cell population) were revealed for each of the two samples respectively. The last column of three plots (3C), (3F) and (3I) showed boxplot of newly constructed variables from EdTAR process for three samples. The x-axis indicated the new variables along the D.I range [0 – 8 ] and y-axis was the boxplot of available values for each variable.

**Figure 4. Statistical model assessment** In order to predict the OCRI, we explored a series of prediction models. Seven models (SVM, RRF, PLR, NNET, KNN, and CART) were shown with three major performance assessment metrics (ROC, sensitivity and specificity), ranging between 0 and 1. Each model were trained on the training data and tested on testing data. Each boxplot showed the distribution each of the three metrics from the model resample assessment (R caret package <http://cran.r-project.org/web/packages/caret/index.html>)

**Figure 5. Oral Cancer Risk Index (OCRI) panel** Based on the model selection and assessment, an oral cancer risk index (OCRI) was established. The process was tested on a hold-off dataset with all three (known) clinical classifications: Normal, OLK, and OSCC. Samples from each class were assessed with the newly proposed ORCI panel. The y-axis showed the ORCI index ranging between 0 and 1, where 0 indicates the lowest risk (of cancer) and 1 indicates the highest risk (of cancer).