**Figure Legends**

**Figure 1.** **Distribution of DNA contents in exfoliative cytology.** (A) Selected cells with abnormally high DI values (>2.3). (B)A scatter plot with y-axis as the area of nucleus and x-axis as DI value. (C) Distribution histogram of DI values of all nuclei. (D) Distribution histogram of DI values of the three cell populations after simulation from normal distribution, diploid cell population (red; µ=1.001, σ=0.19), tetraploid cell population (green; µ=2.002, σ=0.25) and aneuploidy cell population (blue; µ=2.300, σ=0.5). When these three cell populations are merged at the ratio of 0.893:0.092:0.05, a composite distribution histogram (black) can be generated.

**Figure 2. Work flow of expert-guided data transformation and reconstruction (EdTAR).** Starting with DI values as the raw data, EdTAR first identified candidate peaks of cell populations. Diploid cell population was extracted and further filtered if more than one population is detected. The same procedure was applied to extract the tetraploid cell population and thus the aneuploid cell population was isolated. Data of these three cell populations were reconstructed across a wide rage [0 – 8] using the discrete density at each interval. The newly constructed data was used for training the statistical model and calculation of the Oral Cancer Risk Index (OCRI).

**Figure 3. Application of EdTAR in processing data of three samples with pathological diagnosis of normal (A-C), OLK (D-F), and OSCC (G-I).** All density plots have x-axis as DI value and y-axis as density. Panel A, D and G showed density plots before data processing by EdTAR. In Panel A, a major peek with a DI of 0.995 represents the diploid cell population, where another small peaks (DI = 0.594) was a minor population possibly due to image processing. In Panel D, a major peek with a DI of 0.798 represents the diploid cell population (3,590 cells). Other than this peak, four peaks with DI values of 1.25, 1.75, 2.22, and 2.74, were present. In Panel G, a major peek with a DI of 1.02 represents the diploid cell population, and a second peak with a DI of 1.79 represents the tetraploid cell population. Other than these two peaks, three peaks with DI values of 3.25, 3.57, and 3.99 were present, and were believed to represent the aneuploidy cell population. Panel B, E and H corresponding with Panel A, D and G respectively were three plots showing the net results of data processing by EdTAR. Signals of the aneuploidy cell populations were amplified in Panel E and H. Panel C, F and I showed boxplots of newly constructed variables after data processing with EdTAR. The x-axis indicated the new variables along a range of DI [0 – 8] and y-axis the boxplot of available values for each variable.

**Figure 4. Assessment of statistical models.** Seven models (SVM, RRF, PLR, NNET, KNN, and CART) were tested for their performance using three parameters, ROC, sensitivity and specificity. Each model was trained on the training data and tested on the testing data. Each boxplot showed the distribution of these three parameters (R caret package <http://cran.r-project.org/web/packages/caret/index.html>).

**Figure 5. Calculation of Oral Cancer Risk Index (OCRI).** OCRI was calculated for each case with known pathology. The y-axis showed the ORCI between 0 and 1, where 0 indicates the lowest risk of OSCC and 1 indicates the highest risk of OSCC.

**Figure 6. Application of EdTAR in clinical follow-up of one patient (Case 128141).** Exfoliative cytology was performed in April 2008 and a density plot of DI data was generated (A). With EdTAR, positive signals were relatively amplified and an OCRI was calculated as 0.88 (B). Histopathology of biopsy showed mild dysplasia on H&E stained section (C). This patient was regularly followed up in outpatient clinic. A tumor was observed in August 2011. Histopathology of the surgically resected tumor confirmed the diagnosis of squamous cell carcinoma (D).