Running title: Early Detection of Malignant and Pre-malignant using DNA Image Cytometry

Main contribution and objectives:

1. A novel way to process the image data -- ExGCRn (figure 1)
2. Model selection and assessment -- model\_assessment table (table 1)

Abstract

Introduction

**Method and materials**

Oral brushing biopsy method

DNA staining and imaging method

**Expert-guided data cleaning and reconstruction**

For each DNA imaging sample, all available DNA index (D.I.) values were exported from the “software name” [reference here]. With the unknown distribution of the random variable(s), kernel density smoothing function was applied to the mixtures of three potential cell populations (normal, mitotic, and aneuploidy). Gaussian kernel was chosen for the smoothing purpose. With the nonparametric estimation of the probability density function, a mathematical procedure [reference here] was applied to search for local peak(s), which was assumed to represent the mode for each cell population. Based on our biological theory, a few assumptions were introduced. It was assumed that the mean D.I. value for normal cell population was “1c”; for mitotics cell population was “2c”; and the cell population with abnormal dividing (aneuploidy population) was above “2.2c” [reference here].

From here a sequential parsing the D.I. value was implemented. (1) To successfully extract the normal cell population, we searched along the mixed density probability density functions (pdf) and located the peak fell around the vicinity of “1c”. Then, we only used the data on the left of the peak to estimate the mean and standard deviation for the first population. Using the background normal distribution as our assumption with the estimated mean and standard deviation, the normal population was extracted out of the dataset. (2) A similar procedure was applied to extract the mitotic cell population and in the end only data left (if any) belonged to the abnormally dividing population. In case, the dataset only contained normal cell population, no cleaning was applied.

The next step is to reconstruct the “useful” dataset for the statistical modeling. To do so, we need summary statistics (mean and standard deviation) from the first two populations and number of data points in each population. The newly constructed data would normally represent the mixtures of two or three population at a control ratio, which provides the basis for building the following prediction model. (1) If only the normal population was determined, a ratio at: 98:1.5:0.5 was used to reconstruct the final data (2) If both normal and mitotic population were determined, the actual ratio between the two families was used and together consisted 99.5% toward the total leaving unchanged 0.5% for the abnormal population (3) If all three populations were determined, the actual ratio between the two families was used and together consisted 90% toward the total leaving unchanged 10% for the abnormal population.

The last step is to convert the reconstructed D.I. values to sixteen measurements of hidden variables. Now, with the known mixture ratios of population(s) and both the summary statistics and raw data value, we expand the D.I. measurement value on the scale between 0 and 8 (for samples with D.I. values greater than 8, they were converted 8). A mixed density was further consulted and then discretized into 16 bins, the density for each bin was stored as the “measurement” for each hidden variable.

**Building the statistical prediction model and model performance evaluation**

**Prediction performance metrics**

For a particular classifier, the confusion matrix can be summarized in various standard performance metrics to compare MAQC-II model prediction performance.

The Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Total Samples (*TS*) | Actual Positives (*AP*) | Actual Negatives (*AN*) |
| Predicted Positives (*PP*) | True Positives (*TP*) | False Positives (*FP*) |
| Predicted Negatives (*PN*) | False Negatives (*FN*) | True Negatives (*TN*) |

Sensitivity is a measure of actual positives correctly identified as such and the specificity measures the proportion of negatives correctly identified:

*Sensitivity = TP / AP = TP / (TP + FN)*

*Specificity = TN / AN = TN / (TN + FP)*

There is a trade-off between sensitivity and specificity, making models difficult to compare on the basis of these performance metrics. In contrast, such measures as accuracy, the proportion of correct predictions, the ROC curve, and the Matthews Correlation Coefficient (MCC)74, 75, enable a single parameter comparison of performance of binary classification models. The MCC is defined as:

**Some modeling**

**Cleaning DNA index values**

Cleaning the data has been the most challenging part in our analysis. In an aneuploidy sample (128110), there were 2739 D.I. values. Since this is the mixture of three cell populations: normal cells, mitotic cells, and cell with aneuploidy. We would fit the data with mixture model procedure, if the proportion were somewhat balanced. In fact, often time, we observed “normal cell population” took the majority of the density in a density plot. Based on our prior knowledge, we started off with a strong assumption that the first two peaks will represent the means for the normal cells and mitotic cells, where whatever left over would be the “possible” aneuploidy cell population. Therefore, we propose to sequentially (figure ##) strip off the data that belong to the first two populations, with a hope that we can be left off with the signal that we are really interested in. To do so, we leverage the kernel density smoothing technology (reference here…), then we search for the first peak (references here…) along density smooth curve. Here, we brought in another strong assumption, which “no data points” from the following population went off the left side of the first peak. With these assumptions, we estimated the mean and standard deviation using only left side of the first population. These summary statistics were saved for future use. To strip off the left side of the data was quite straightforward; however stripping the data on the right side of the peak needed specially handling.

**Building the predication model**

After cleaning all the dataset, we went ahead and tested out predication models (reference here..)