Method and materials

**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..)

Major contributions/achievement

**Parse the data to make “minor” category visible**

**Comparison between parameters ..**

**Building the predication model based on “normal” and “oscc”**

**Building the predication model on all three cases**

**Automated data analysis and machine learning tactic**