**Ex**pert **G**uided Data **C**leaning and **R**econstructio**n** (**ExGCRn**) procedure

***Expert input:***

Param1: Threshold for aneuploidy D.I.

Param2: Mean D.I. value for mitotic cell population

Param3: Standard deviation for mitotic and aneuploidy cell populations

Param4: Ratio of normal cell (both normal and mitotic) vs. aneuploidy when three populations observed

Param5: Ratio of normal cell (both normal and mitotic) vs. aneuploidy when no aneuploidy population observed

Param6: Ratio of normal cell vs. mitotic population when no aneuploidy population observed

***Preparation phase***

Read in raw DNA index values (D.I. value)

Determine how many cell populations exist in the exfoliated cell population

**If** three populations determined

Use expert guided parameter set (param4)

**Else If** no aneuploidy population observed

**If** no mitotic population observed

Use expert guided parameter set (param2, param3, param6)

**Else**

Use expert guided parameter set (param2, param3, param5)

***Stripping the normal population***

Step 1 Derive the probability density function with kernel density estimation smoothing

Step 2 Determine the mode of the normal cell population

Step 3 (a) Use the data to the left side of the mode

(b) Estimate the mean and standard deviation of the normal population

Step 4 Remove the data point to the left the mode

Step 5 (a) Find the symmetric subset of the data on the right of the mode

(b) Break the data on the right into equal **n** parts according to the KDE density

***Foreach*** interval

Derive the theoretical number of data points

Randomly select candidate data points to be filtered

Recover missing data points due to rounding errors

End of ***foreach*** loop

Step 6 --> go back to step 1 and evaluate the stripping efficiency

***If***the first left most mode is less than the threshold

Redo step 2 – step 6

***Else***

Finish stripping the normal population

Compute the mean and standard deviation from the stripped first population only

Store number of data points filtered

***Stripping the mitotic population***

Step 1 Follow the same “step 1 – step 6” from the stripping the normal population

Step 2 Stop the cleaning ***if*** (1) the first left most mode is less than the threshold ***or*** (2) three rounds have been reached

Step 3

Finish stripping the mitotic population

Compute the mean and standard deviation from the stripped second population only

Store number of data points filtered

***Data reconstruction***

The reconstruction was handled according to the number of populations determined in the cleaning steps

Step 1 If three populations were determined, a cap of “8” was applied for the D.I. values

Step 2 Parameters either from the **expert input** or derived from the data itself were used

Step 3 A heuristic cumulated density function was derived and used to guide the “random sampling” of the data points

Step 4 The dataset was stretched/binned between [0, 8] on the D.I. axis. The density within each bin will be recorded as the measurement of that “variable”. Any missing values were replaced with 0.00001