**Department of Computing and Information Systems**

**COMP 90016**

Workshop 8

The lectures introduced a different way to analyse structural variants, by looking at anomalously paired-end reads. The concept is to set an expected range of insert sizes to come from the sequencing machine and then to analyse read pairs that do not fall into this range for potential structural variants, such as duplications, deletions, insertions, or inter-chromosomal events.

In this workshop we are going to do the groundwork for this approach by establishing the distribution of insert sizes and what should be the normal range of values.

The *Pysam* library for *Python* has great functionality to open a file of aligned reads (in the BAM format) and extract their insert sizes (for paired reads only).

Write a program to do the following:

1. Open the BAM file *paired\_reads.bam.*
2. Have it write the first 10 records to the command line, and run your script.
   1. Can you identify the fields indicating the pairing of reads?
   2. Are all reads paired as expected?
3. Extend your script to parse the first 10,000 lines and keep track of the insert sizes. Only collect values for paired reads with a meaningful insert size (reads that are actually paired and aligned to the same chromosome).   
   Note that each read pair shows up twice in the data: Once for the first and then again for the second read. Record the insert size only once.  
   After parsing 10,000 lines, use the *numpy.mean()* and *numpy.std()* functions to report the median insert size and standard deviation to the command line.   
   You can also visualise the data with a histogram or density plot. Does the distribution look like a normal distribution?  
   Note that the mean value and standard deviation are subject to influence of outliers in the data. It might be a better approach to run the analysis on a subset of the data in the 20-80% percentile. Also, the median might be more robust.
4. Parse the entire file as above (only considering pairs once etc.), and report any value that is outside of the mean +- twice the standard deviation.   
   For a normal distribution, we would expect 95% of the data to fall within the specified range.
   1. Does the data conform to this expectation?
   2. What does this say about the existence and abundance of structural variants in the data?