**Onco-Wiki Quality Control**

1. **Reliability of Data**

Since we are using real data and many publications fail to sequence healthy patient samples, it is important to gauge variability in the sample groups you have and in the Healthy Dataset. Variations in sequencing technology, sample acquisition and RNA isolation protocol could have downstream impacts on the significance of particular genes.

1. **Principal Component Analysis**

Principal Component Analysis plots clearly portrays the variance of sample grouped data into clusters by taking two the largest sources of variation in gene expression as the two axes on the graph.

Go to Toolbox | Expression Analysis | Set Up Experiment. Select your recently downloaded Healthy dataset first and all of your samples that you have sequenced. If you have more than 2 samples in total, do an unpaired multi-group comparison and group your samples accordingly. Ensure that the Healthy samples has its own group. Save this comparison file.

Go to Toolbox | Expression Analysis | Quality Control | Principal Component Analysis and use your recently saved file. Add labels or a figure legend and export this plot as a .pdf or screenshot this graph.

1. **Scatter Plots and Box Plots**

Scatter and boxplots between 2 or more samples respectively can help to illustrate variation between samples more concisely.

1. **Further Plots**

It is possible to do further statistical analysis plots on CLC Genomics Workbench such as Volcano and MA plots. Volcano plots help to illustrate significant genes through p-value significance. MA plots help to determine if normalisation or transformation is required on your data.

Normalisation methods including scaling, quantiles and total adjustments. These are applied to reduce technical variation in your fold changes. You can recreate your original graphs using these new normalised expression values.