Distances Comparison Results (Gene expr.-TFs-Pathways)

Gene expression distances vs TFs distances

The initially computed distances at gene expression level **based on the ranked matrix** and transcription factor activity level **based on the ranked matrix** are plotted against each other for the three cell lines in order to uncover possible correlations between them. The results are the following.

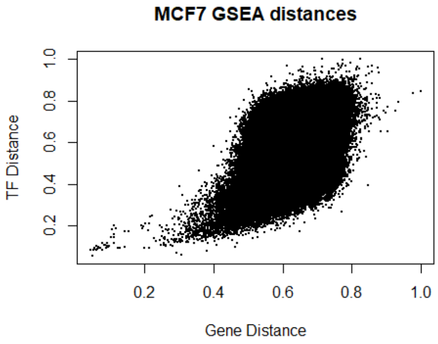

Figure 1 : MCF7 Cell Line, correlation: 0.41405

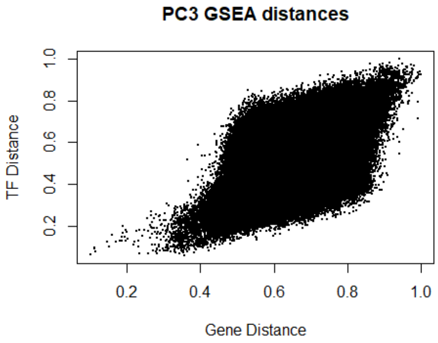

Figure 2 : PC3 Cell Line, correlation: 0.46024

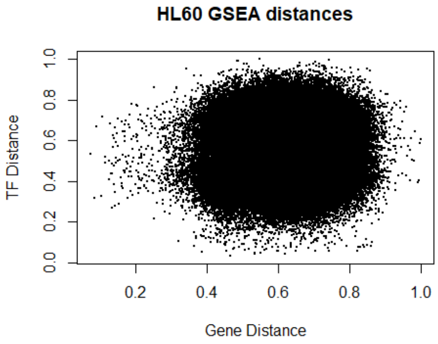

Figure 3 : HL60 Cell Line, correlation: 0.06287

As it can be observed, the correlation in the first two cell lines (MCF7 and PC3) between the two pairwise distances of the compounds exists and is around 0.4 to 0.46. At the HL60 cell line there seems to be no correlation between the two distance metrics.

HL60 further investigation

The fact that the correlation between the gene expression distance **based on the ranked matrix** and the transcription factor activity distance **based on the ranked matrix** at the HL60 cell line is near to zero, indicated the need of a further analysis. It is possible that this result stems from the existence of two different arrays in the cell line (HG-U133A and HT-HG-U133A), while the other cell lines had one array each. So, the distances are plotted for each one of the arrays that are present in HL60 cell line and the plots are shown below.

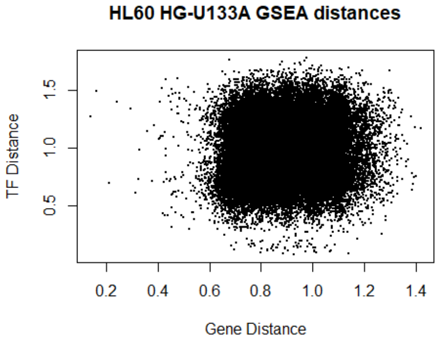

Figure 4 : HL60 Cell Line, HG-U133A Array, correlation: 0.1166783

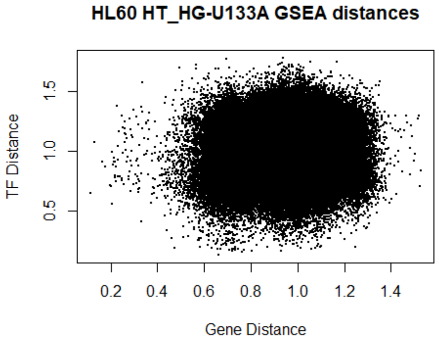

Figure 5 : HL60 Cell Line, HT-HG-U133A Array, correlation: 0.09012696

Again, there seems to be no apparent correlation between the pairwise distances according to the two metrics even when separated to each array.

One possible explanation can be that the HL60 cell line is a leukemia cell line and its cells are heterogeneous and polydynamic. This may lead to them being expressed differentially and as a result, the gene expression distance metric is unreliable and the correlations are near to zero.

TFs distances vs Pathway distances

As a next step, the pathways affected by each compound in each cell line, based on the below mentioned transcription factor activity, were calculated and then their distances were computed, using GSEA. These findings were then compared to the distances according to the transcription factor activity of the compounds **based on the gene ranked matrix**. The following plots and correlations are this comparison's results.

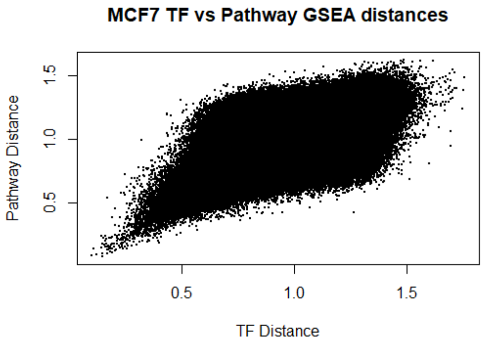

Figure 6 : MCF7 Cell Line, correlation: 0.5527844

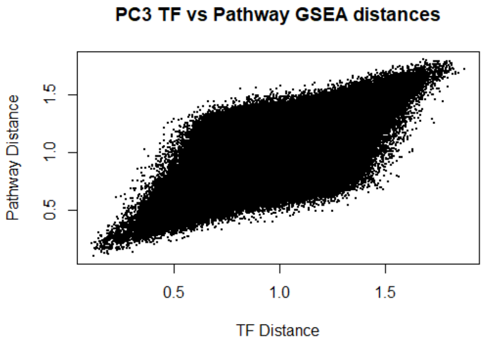

Figure 7 : PC3 Cell Line, correlation: 0.6497039

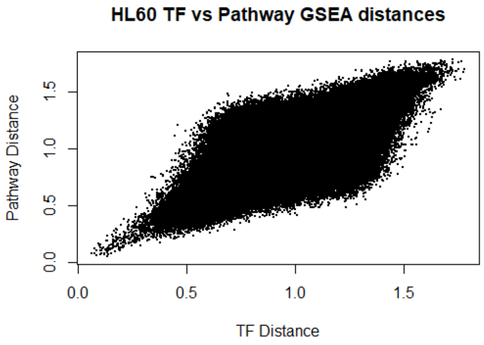

Figure 8 : HL60 Cell Line, correlation: 0.6367222

According to the above results, the correlation between the TF distances and the ones derived from the pathways are correlated at all of the three cell lines, with high rates (above 0.6). This indicates a strong relationship between the TF distances and the pathway distances.

Pathway computation techniques

There are two ways that the pathways can be derived. On one hand, one can use the ranked gene expression matrix provided by Iorio's data. On the other hand, the log(FC) matrix can be derived from the raw experimental data. The difference between these two is that a different pipeline is followed for the construction of each dataset. On this ground, a comparison between the pathway pairwise distances that are derived from each pipeline of gene expression (ranked gene matrix and log(FC)) is meaningful in order to get an overview of the two methods and their results. The plots that are constructed by the two distances for each cell line are shown below.

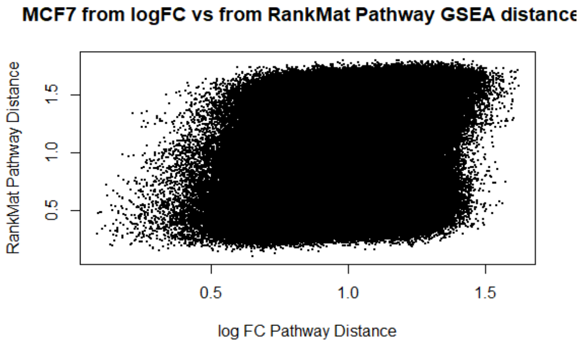

Figure 9 : MCF7 Cell Line, correlation: 0.2568513

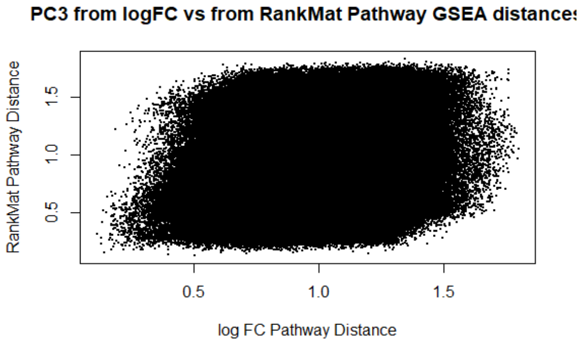

Figure 10 : PC3 Cell Line, correlation: 0.2048015

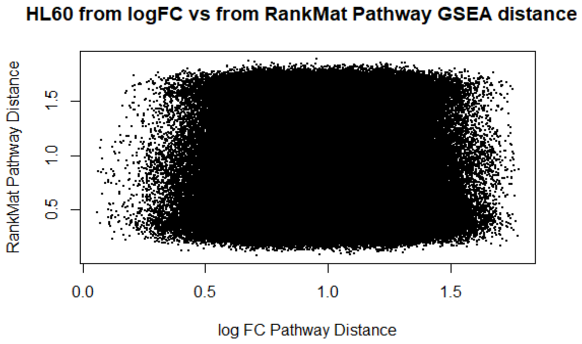

Figure 11 : HL60 Cell Line, correlation: 0.01157822

From the above results, it is obvious that in pathway level no correlation exists between the two alternatives of gene expression computation. Because of this fact, the gene expression vs pathway distance will be calculated two times, one for each gene expression calculation method.

Gene expression distances vs pathway distances

1. Gene expression from log(FC)

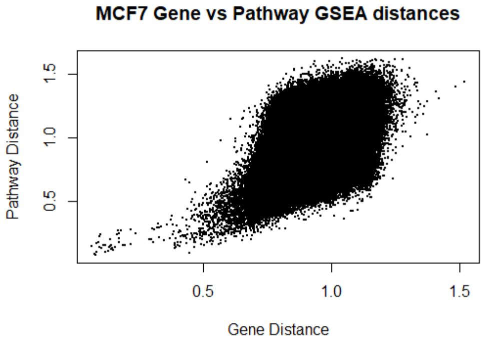

Figure 12 : MCF7 Cell Line, correlation: 0.3923077

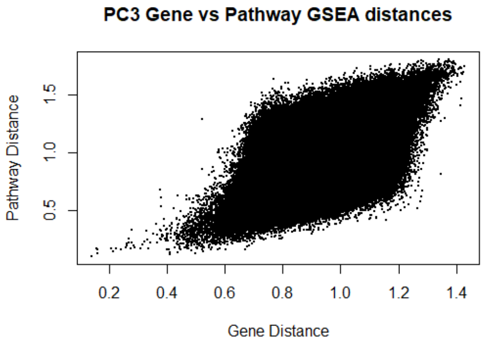

Figure 13 : PC3 Cell Line, correlation: 0.4742973

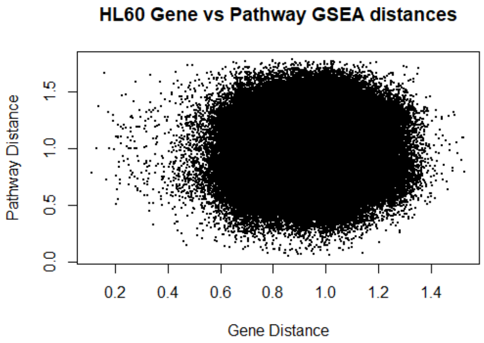

Figure 14 : HL60 Cell Line, correlation: 0.05327427

According to the results, a correlation is uncovered in the MCF7 and PC3 cell lines, while no correlation seems to exist in the HL60 cell lines between the gene expression distances according to the log(FC) and the pathways derived from them.

2. Gene expression from ranked matrix

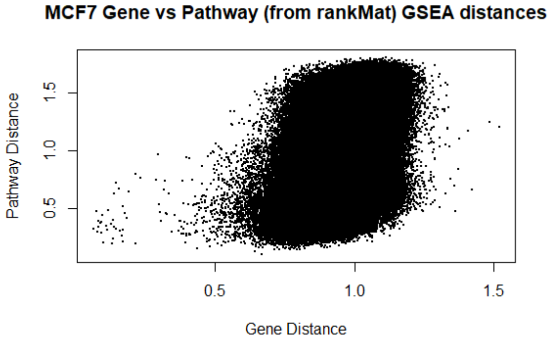

Figure 15 : MCF7 Cell Line, correlation: 0.5246859

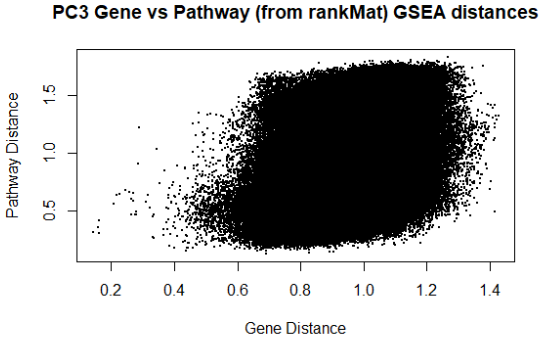

Figure 16 : PC3 Cell Line, correlation: 0.4293407

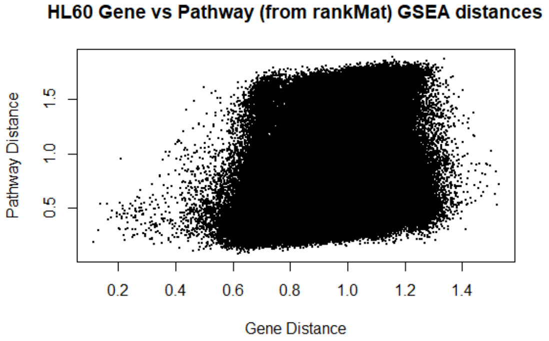

Figure 17 : HL60 Cell Line, correlation: 0.4198651

These results exhibit a stronger correlation in MCF7 cell line, a little weaker in PC3 cell line and a correlation of 0.42 in HL60 cell line, whose distances were previously uncorrelated.

The difference between the correlations with each method stems from the gene expression calculation model and this will have to be further investigated and finally only one approach of gene expression calculation will be selected to be used in the study.