Gene Co-Expression Network of Whole Cancer Cell Lines

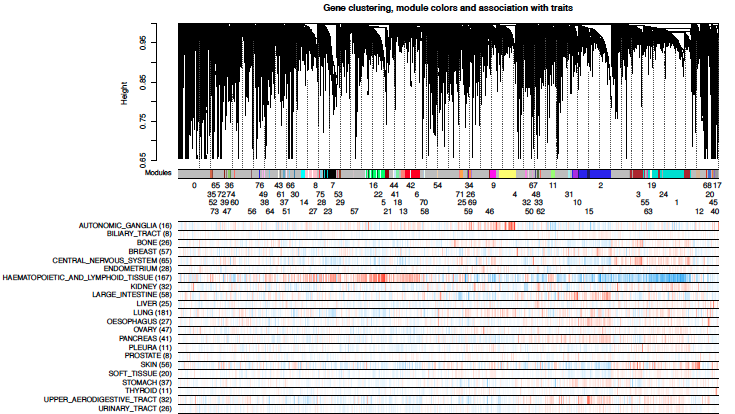


Fig.1 Gene clustering, module colors, and association with traits

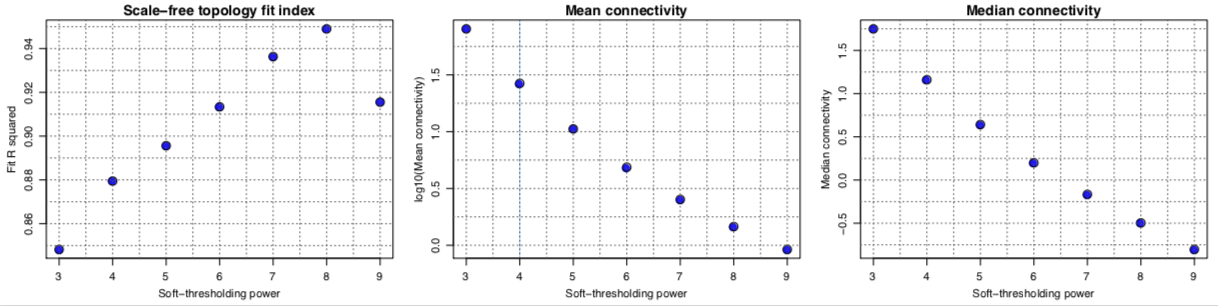
This analysis I had developed the R code for preprocessing RNA-Seq data and network analysis, which I used the knowledge from the fundamental of WGCNA (Langfelder and Horvath 2008). The script pre-processes the raw counts by filtering out low-expressed features, running an approximate variance-stabilizing transformation, and calculating observation whose purpose is to downweigh outliers. Also, the script creates rudimentary sample traits by inferring the type of cancer from sample names and creating binary indicator variables for cancers that have at least 5 samples. And the network analysis includes scale free topology analysis, network construction, plot gene dendrogram, module colors and association with cancers, and functional enrichment analysis of the modules genes (GO terms).

Fig.2 Scale-free topology analysis

The Scale-free topology fit is shown in Figure 2. The scale-free topology plot shows good scale-free topology fit index at power 4 (0.88), and the mean and median connectives are in a range that I would like to see, log10(connectivity) between 1 and 1.5. From the result, I could run the analysis with the powers anywhere between 3 and 6 and I would get similar (though of course not the same) modules, at least the large ones. In case of the mean connectivity, for each power, connectivities were calculated for each gene and the mean over all genes is shown in the figure. The median connectivity, for each power, connectivities were calculated for each gene and the median over all genes is shown in the figure as well.

The gene modules were detected by using average linkage hierarchical clustering coupled with a gene dissimilarity measure to define a dendrogram (cluster tree) of the network. Moreover, each row corresponds to an indicator of cancer type: the indicator is 1 if the sample belongs to the cancer type and 0 if it does not. Then correlate the indicator variables with (variance stabilised) gene expression values. Each row then depicts the correlations with blue colour meaning negative correlations and red colour positive correlations. In other words, genes (and modules) with a lot of red are up in the cancer, and the ones with a lot of blue are down in that cancer. The result is shown in Figure 1.