TCGA coexpression network TODO

* Familiarize self with TCGA data
* Find analogous RNAseq data
* Read about TCGA file naming
  + Do I have any biological/ technical replicates? ..No
* Normalize microarray..Actually, I don’t need to.
* Hierarchically cluster samples to see if there are subgroups
* Normalize RNAseq
* R project tutorial
* Outline pipeline
* Compute correlation statistic
  + Pearson
  + Spearman
  + Biweight midcorrelation bicor
  + Euclidian distance
  + Canonical correlation (I may need exon-level counts for this)
  + Distance correlation
* Create network soft threshold
  + Hard thresholding
  + Soft thresholding
    - Signed – aij = |cor(xi, xj)|^B
    - Unsigned - aij = | 0.5 + 0.5 \* cor(xi, xj)|^B
* Calculate difference of networks
* Visualize networks
  + Cytoscape
* Determine metrics for coexpression network similarity.
  + Mutual information
  + Correlation
  + Model based indices
  + Topological overlap measure (number of shared friends)
  + Module preservation statistics\
  + Person correlation (R value) of scale-freeness log(connectivity aka soft thresholded degree) vs log(frequency)
    - Choose a power so that the resulting network is scale-free
  + Number of hub genes
    - Are they the same hub genes?
  + PCC distribution (should be normal about 0)
  + PCC between samples box plots (one for microarray one for RNASeq like Arabidopsis thaliana paper by Giorgi et al.)
  + Normalizing graphs
    - Can I find a way to identify differentially expressed genes across the two technologies? I’ll have to identify distributions of genes and normalize for each gene I think. This will give me the set of genes that
    - Consider the following cases for gene A and B:
    - A and B are differentially expressed across technologies
      * A correlates with B in micro array
        + A correlates with B in RNASeq

Graphs are similar

* + - * A and B do not correlate
        + Graphs may not be similar