Gene Expression Analysis

With MATLAB

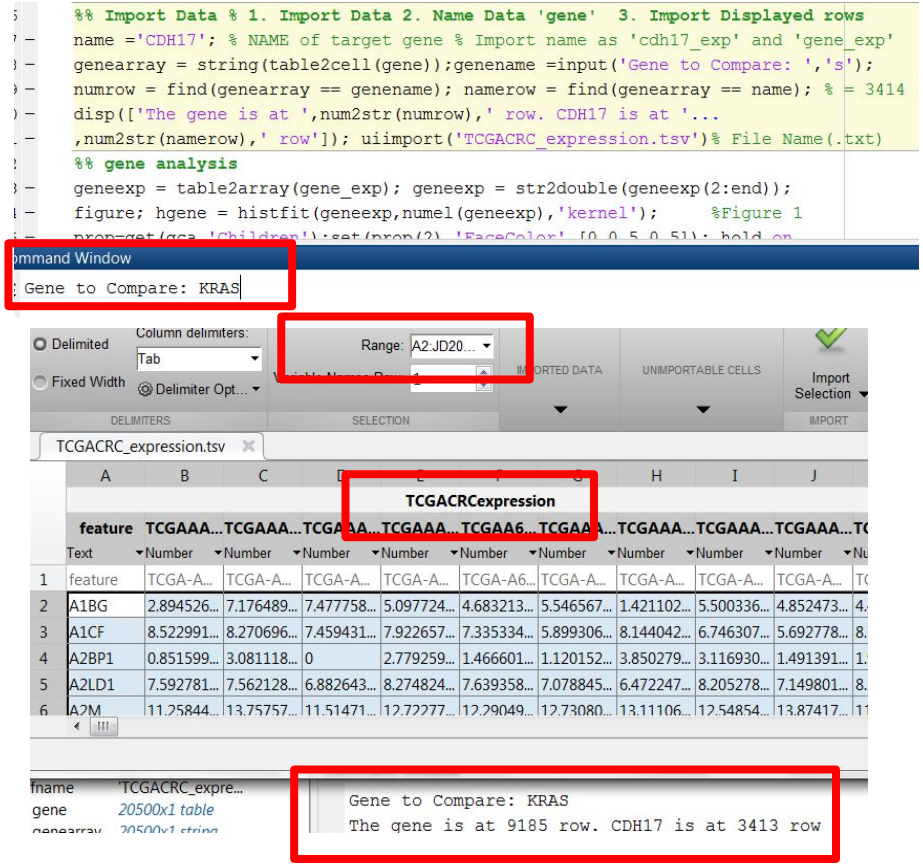
Data Sources

- Colorectal Cancer Subtyping Consortium (Synapse)
  - https://www.synapse.org/#!Synapse:syn2634724
  - geneExpression/TCGA/TCGACRC_expression.tsv
  - TCGA dataset (click here for link)
  - Consist of normalised gene expression data by patients
  - Referenced in (Guinney, Justin, et al. "The consensus molecular subtypes of colorectal cancer." *Nature medicine* (2015).) (shown by Lum) “Tumor purity analysis. We obtained the tumor purity estimation of C in the TCGA data set as defined ...’

Matlab functions

- Using read and find function to import desired gene dataset for comparison
  - Can be fully automated if done on a laptop with enough memory
- Creating Histogram based on expression data
- Using histogram normal distribution fit to estimate overall expression derived from the gene
  - Compare the shift in the distribution
  - Available in different modes of fitting
- Calculate correlation coefficient of expression across patient samples
- Scatter plot with linear fit to compare expression
  - Calculate linear regression

Directions



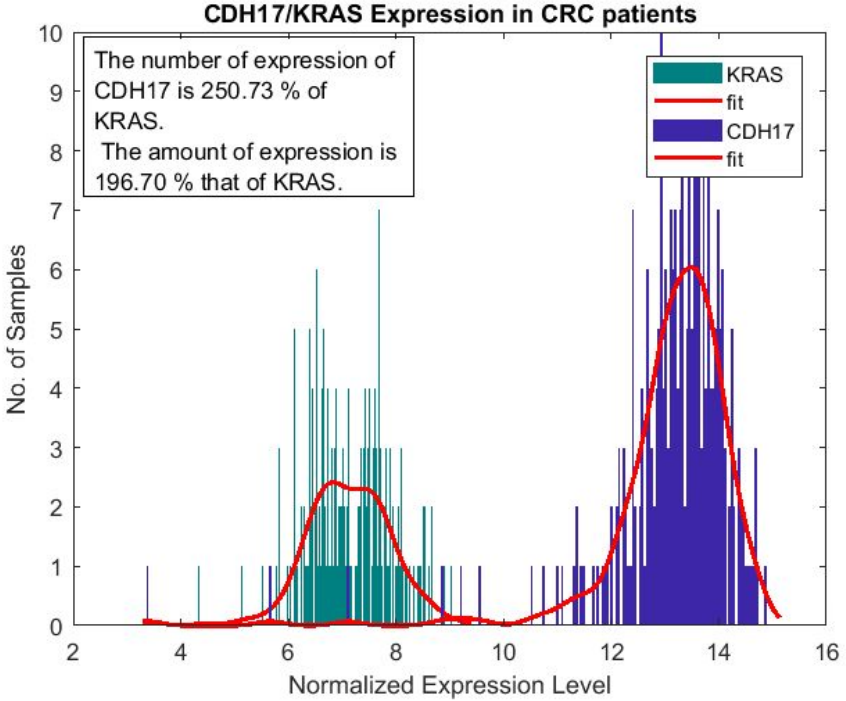
1. Input gene name desired to compare with target gene CDH17

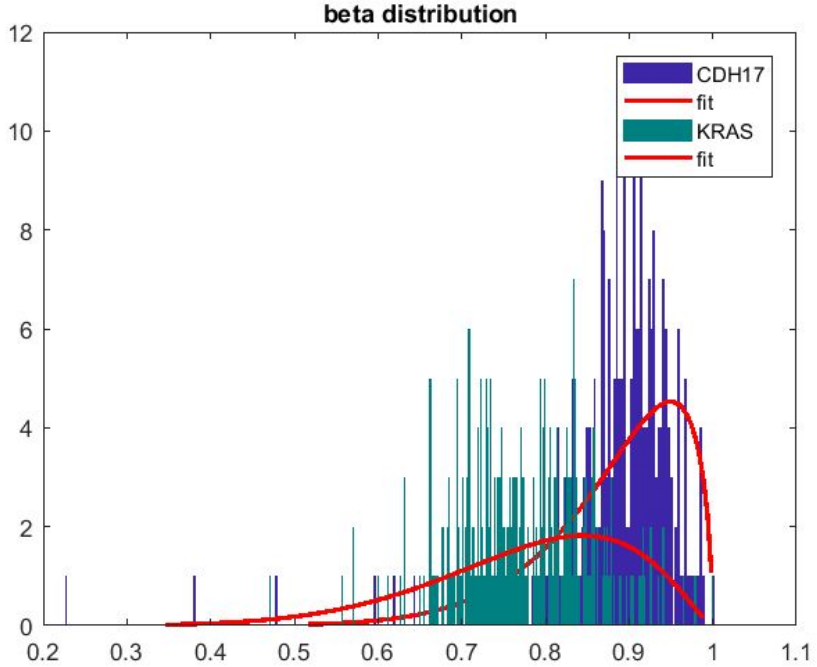
2. Import Data
   - Use Step 1. to find and input range
     - CDH17(give name ‘cdh17_exp’)
     - Targetname (give name ‘gene_exp’)
     - Import both

3. Run the program
   - Enter to terminate after run

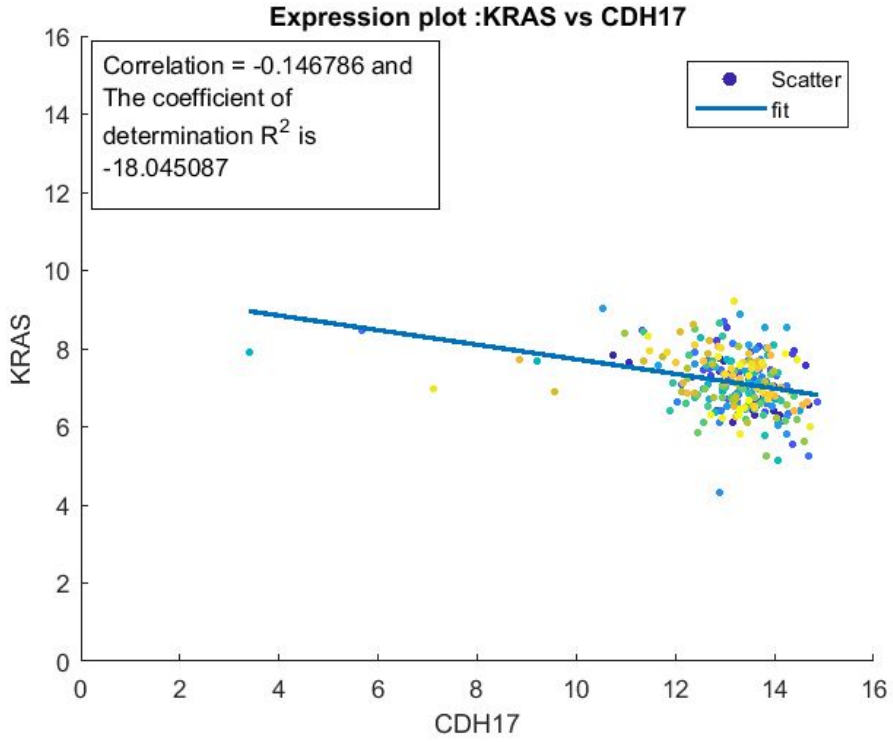
Results

- Runtime: >10 seconds



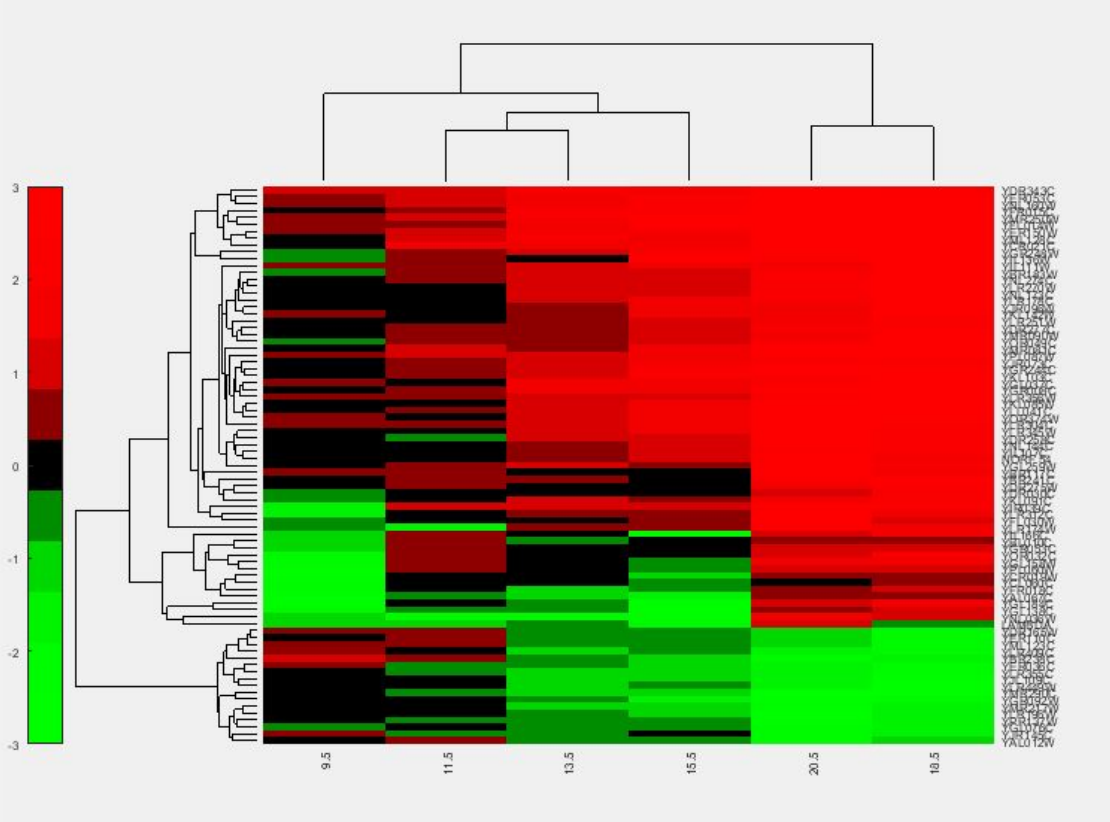


Normalised to 0 to 1



- Correlation coefficient between 2 data set
- Coefficient of Determination R^2

Other Analysis -Matlab Hierarchical Clustering



- Example results regarding yeast expression
- Datasource: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28

(Required: Expression over time)

- Matlab Example:
- Click Here for link
- Source:

(Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, *278 (5338)*, 680–686. PMID: 9381177.)

Function - D'haeseleer, Patrik. "How does gene expression clustering work?."

- 'The goal of clustering is to subdivide a set of items (in our case, genes) in such a way that similar items fall into the same cluster, whereas dissimilar items fall in different clusters'
  (D'haeseleer, Patrik. "How does gene expression clustering work?." *Nature biotechnology* 23.12 (2005): 1499.)
- Employs different algorithms to subdivide datasets using different similarity measures

Table 1 Gene expression similarity measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (\mathbf{e}_f - \mathbf{e}_g)' \Sigma^{-1} (\mathbf{e}_f - \mathbf{e}_g)$, where Σ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace e_{gc} with the rank of e_{gc} within the expression values of gene g across all conditions $c = 1 \ldots C$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.