



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

```
%% Import Data % 1. Import Data 2. Name Data 'gene' 3. Import Displayed rows
name = 'CDH17'; % NAME of target gene % Import name as 'cdh17_exp' and 'gene_exp'
genearray = string(table2cell(gene)); genename = input('Gene to Compare: ','s');
numrow = find(genearray == genename); namerow = find(genearray == name); % = 3414
disp(['The gene is at ', num2str(numrow), ' row. CDH17 is at '...
, num2str(namerow), ' row']); uiimport('TCGACRC_expression.tsv') % File Name (.txt)
%% gene analysis
geneexp = table2array(gene_exp); geneexp = str2double(geneexp(2:end));
figure; hgene = histfit(geneexp, numel(geneexp), 'kernel'); %Figure 1
prop = get(gca, 'Children'); set(prop(2), 'FaceColor', [0.0 5.0 51]); hold on
```

Command Window

Gene to Compare: KRAS

Delimited		Column delimiters:	Range: A2:JD20...		IMPORTED DATA		UNIMPORTABLE CELLS		Import Selection	
Fixed Width		Delimiter Opt...								
TCGACRC_expression.tsv										
TCGACRCexpression										
feature	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...	TCGAAA...
Text	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
1 feature	TCGA-A...	TCGA-A...	TCGA-A...	TCGA-A...	TCGA-A6...	TCGA-A...	TCGA-A...	TCGA-A...	TCGA-A...	TCGA-A...
2 A1BG	2.894526...	7.176489...	7.477758...	5.097724...	4.683213...	5.546567...	1.421102...	5.500336...	4.852473...	4.852473...
3 A1CF	8.522991...	8.270696...	7.459431...	7.922657...	7.335334...	5.899306...	8.144042...	6.746307...	5.692778...	5.692778...
4 A2BP1	0.851599...	3.081118...	0	2.779259...	1.466601...	1.120152...	3.850279...	3.116930...	1.491391...	1.491391...
5 A2LD1	7.592781...	7.562128...	6.882643...	8.274824...	7.639358...	7.078845...	6.472247...	8.205278...	7.149801...	7.149801...
6 A2M	11.25844...	13.75757...	11.51471...	12.72277...	12.29049...	12.73080...	13.11106...	12.54854...	13.87417...	13.87417...
fname TCGACRC_expre...										
gene 20500x1 table										
genearray 20500x1 string										
Gene to Compare: KRAS										
The gene is at 9185 row. CDH17 is at 3413 row										

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

2. Import Data

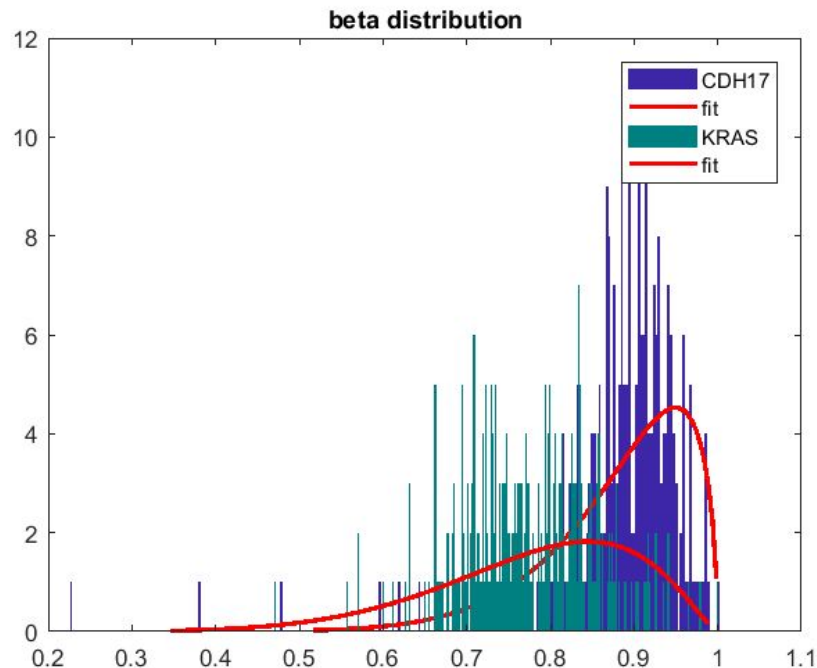
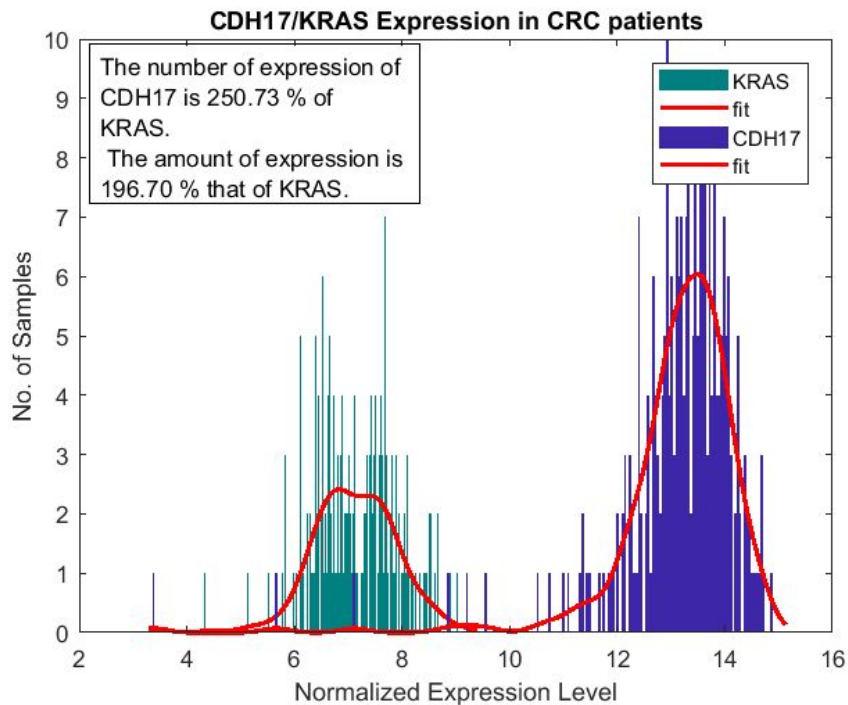
- Use Step 1. to find and input range
 - CDH17(give name 'cdh17_exp')
 - Targetname (give name 'gene_exp')
 - Import both

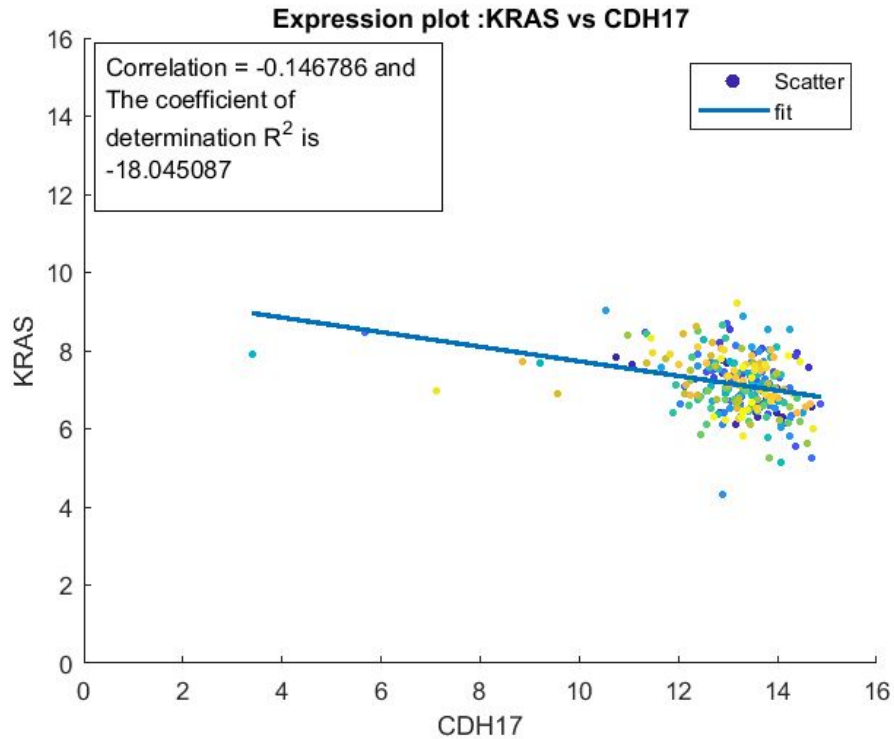
3. Run the program

- Enter to terminate after run

Results

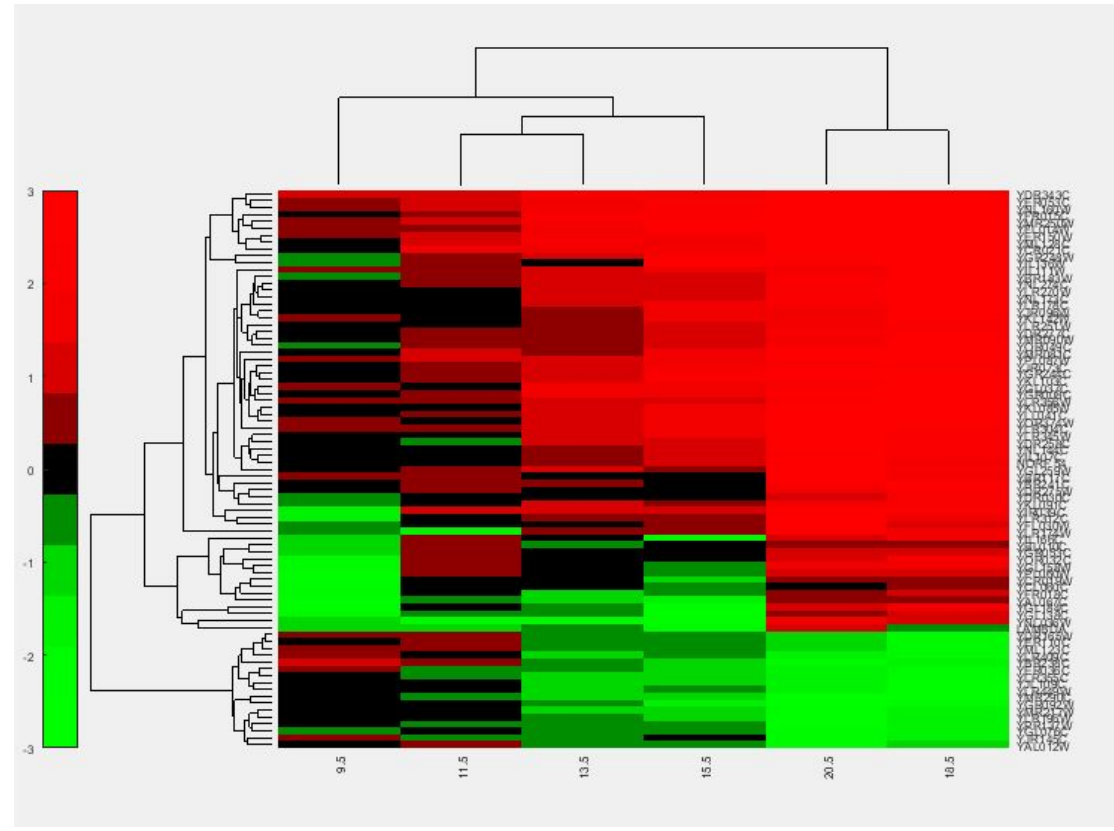
- Runtime: >10 seconds





- 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 e_{fc} - e_{gc} $
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 $c = 1 \dots C$
Absolute or squared correlation	$d_{fg} = 1 - r_{fg} $ 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 .