Noah M. Jorgenson

Bioinformatics CS6643

Lab 9

|  |  |  |
| --- | --- | --- |
| 1. [**Series GSE5847**](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5847) |  | [Query DataSets for GSE5847](http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds&term=GSE5847%5BAccession%5D&cmd=search) |

Status Public on Sep 30, 2007 Title Tumor and stroma from breast by LCM Organism [Homo sapiens](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606) Experiment type Expression profiling by array  
Summary Tumor epithelium and surrounding stromal cells were isolated using laser capture microdissection of human breast cancer to examine differences in gene expression based on tissue types from inflammatory and non-inflammatory breast cancer  
Keywords: LCM  
  Overall design We applied LCM to obtain samples enriched in tumor epithelium and stroma from 15 IBC and 35 non-IBC cases to study the relative contribution of each component to the IBC phenotype and to patient survival.

The overall experiment we will be preforming is based on a MATLAB demonstration utilizing the Bioinformatics Toolbox. The description they gave at the demo website is as follows:

*In this demonstration, you will retrieve the GSE5847 data set from GEO database, and perform statistical testing on the data. GEO Series GSE5847 contains experimental data from a gene expression study of tumor stroma and epithelium cells from 15 inflammatory breast cancer (IBC) cases and 35 non-inflammatory breast cancer cases (Boersma et al. 2008).*

1. When counting how many genes were cutoff during our FDR analysis I noticed that since the q-value versus p-value graph starts around q-value ~= 0.18 we should instead count maybe around the area around 0.30 or those genes with q-values less than 0.40-0.50 (looking at the graph and the results of various numbers from 0.18 to 0.30 this seems resonable). The adjusted FDR p-values too are much too large to be in the range below the cutoff of 0.05 and exist only for p-values up around 0.5.
2. I encounted lots of issues with running the lab for the final section where we had to perform the gene ontology analysis. So I had to manually obtain the GPL data files and reference then instead of grabbing them on the fly in my script with the geosoftread function which can read in the soft data from a txt file. After that I tried to simply update my old lab9.m script to use the gene ontology sections from your solution script but that gave me a rather cryptic error about subscripts being incorrect. So, I opted to simply run the solution script on it’s own and interpret the results.

The resulting GO Terms:

GO Term p-value counts definition

GO:0034386 0.00167 1 / 3 Catalysis of the reaction: 4-aminobutanoate + 2-o...

GO:0034387 0.00167 1 / 3 Catalysis of the reaction: 4-aminobutanoate + pyr...

GO:0003867 0.00445 1 / 8 Catalysis of the reaction: 4-aminobutanoate + ami...

GO:0004621 0.00445 1 / 8 Catalysis of the reaction: glycoprotein phosphati...

GO:0047298 0.00445 1 / 8 Catalysis of the reaction: (S)-3-amino-2-methylpr...

GO:0005364 0.00556 1 / 10 Catalysis of the transfer of a solute or solutes ...

GO:0005402 0.00556 1 / 10 Catalysis of the transfer of a solute or solutes ...

GO:0008506 0.00556 1 / 10 Catalysis of the transfer of a solute or solutes ...

GO:0009679 0.00556 1 / 10 Catalysis of the transfer of a solute or solutes ...

GO:0015518 0.00556 1 / 10 Catalysis of the transfer of a solute or solutes ...

I noticed that none of the terms are really ‘enriched,’ I think enriched means they appear in a significant amount but these seem to be fairly spread out and only appear once. Nevertheless the expressed genes were:

