UCSD Bioinformatatics Bootcamp Homework 1

Olga Botvinnik

September 20, 2012

Note: for this homework, I used `pandas` version 0.8.1, available here: `http://pandas.pydata.org/getpandas.html`. The version in the Enthought distribution of python, pandas 0.7.3, does not properly account for duplicately named rows, as found in this dataset. I downloaded the source tarball and used the command `python setup.py build ; sudo python setup.py install` to install.

A question I had on this dataset: What are the "call"s of *A*, *P*, or *M*?

A **How many distinct genes are represented in the data set?** While the description of the dataset states there are 6416 genes, there are 7229 rows and 5002 distinct genes, as determined by `len(df.index.unique())`. Where `df` is the imported dataset in dataframe format. *Note: if using Ryan's dataset and the same function, you get 5001. If you use Unix scripting, you get another value...*

```
$ cut -f 1 data_set_HL60_U937_NB4_Jurkat.txt | uniq | wc -l
  6572
```

and yet another if you use Unix scripting on the csv file:

```
$ cut -f 1 -d, data_set_HL60_U937_NB4_Jurkat.csv | uniq | wc -l
  6424
```

Both of these *NIX methods include the first row. so there is one fewer genes than the value given.

So I'm not sure which is my final answer, since different "unique" functions give different answers.

B **Which two time points are the most highly correlated for each cell type?**
The most correlated samples within HL60 are: HL60_0_hrs and HL60_4_hrs (correl = 0.9662).
The most correlated samples within U937 are: U937_0.5_hrs and U937_4_hrs (correl = 0.9647).
The most correlated samples within NB4 are: NB4_24_hrs and NB4_48_hrs (correl = 0.9661).
The most correlated samples within Jurkat are: Jurkat_0.5_hrs and Jurkat_4_hrs (correl = 0.9789).

C **Which two cell types are the most similar?**
The sample types that are most highly correlated are Jurkat and NB4 (sum-squared of correl matrix: 13.3490).

D It is often useful to know which genes change very little across samples for the sake of normalization or calibration. **Based on this data set, what are ten good candidates for genes to use to calibrate machinery or analyses across all these samples?**
The least variant genes are:
ESTs, Highly similar to EUKARYOTIC INITIATION FACTOR 1A [Homo sapiens] (variance = 2.7353)
ESTs (variance = 2.8824)
No cluster in current Unigene and no Genbank entry for M95586 (qualifier M95586_r_i) (variance = 3.6324)
PCCB Propionyl Coenzyme A carboxylase, beta polypeptide (variance = 3.7794) ESTs (variance = 4.1838)
GIPR Gastric inhibitory polypeptide receptor (variance = 4.2353)
Homeodomain protein (Prox 1) mRNA (variance = 4.2647)
GRIA2 Glutamate receptor, ionotropic, AMPA 2 (variance = 4.2794)
ESTs (variance = 4.3603)
L-arginine:glycine amidinotransferase (variance = 4.4044)

E **Do any genes show two-fold higher expression at 24 hours versus 0 hours for all four cell types? If so, which ones?**
There are 466 genes that are two-fold differentially expressed in all sample types.

F **Which genes are differentially regulated (at least two-fold higher or lower) in HL60 cells as compared to U937 cells at 0 hours?**
There are 3708 differentially expressed genes between HL60 vs U937 at time _0_hrs, written to `diff_expr_HL60_vs_U937_at_t=_0_hrs.txt`.

G Take the list of Gene Accession codes from (F), and run them through the DAVID ontology analyzer. (at `http://david.abcc.ncifcrf.gov/summary.jsp`. These are GenBank Accession codes.) **Are there any enriched ontology terms?**
Not all the GenBank Accession IDs mapped. The six most enriched categories are:

GO category	Enrichment	Description
GOTERM_CC_1	95.5%	Cellular Component, broadest terms of GO tree
GOTERM_CC_ALL	95.5%	Cellular Component, all components of GO tree
GOTERM_MF_1	95.2%	Molecular Function, broadest terms of GO tree
GOTERM_MF_ALL	95.2%	Molecular Function, all terms of GO tree
GOTERM_BP_ALL	94.5%	Biological process, all terms of GO tree
GOTERM_MF_2	94.5%	Molecular Function, more specific than category 1

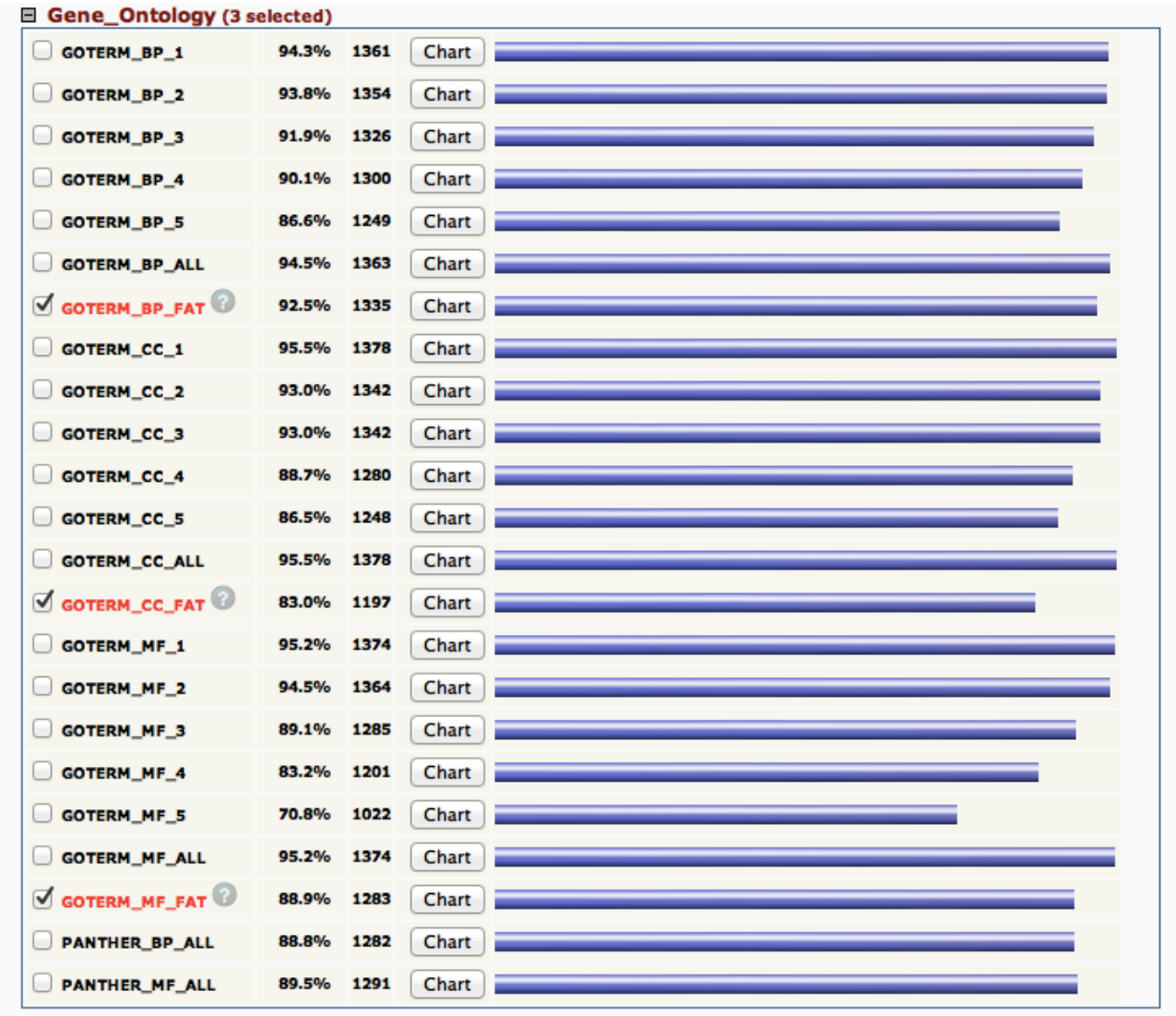

Figure 1: Screen shot of the enriched GO terms from the DAVID website, using the terms found in F.