Supplementary File 1 - Guide to offline phenoclustering

1. Download and install the CluTo Software (free to download & use) from:
http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/cluto-2.1.1.tar.gz
A detailed manual for CluTo can be found here:
http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/manual.pdf

2. Download and install the doc2mat program (free to download & use) from:
http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/doc2mat-1.0.tar.gz

3.
a.) Send an eMail to info@metapark.ch and request a username and password to the PhenomicDB download site. Download is free of charge and no further registration is necessary.

b.) Using the login credentials, download phenotypes in textual form along with associations to Genotypes via Entrez Gene ID from:
http://www.phenomicdb.de/downloads.asp

c.) Format the downloaded phenotypes in the following way:
All phenotypes should be combined into a single textfile, in which each phenotype should begin in a new line with one numeric identifier (e.g. identifying the associated gene) as the first "word" of that line.
Example:

```
1234 This is a phenotype.
2345 This is another phenotype.
3456 This is yet another phenotype.
```

4. Download the (stopwords) list of the most common words in PubMed from:
http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp

5. Run doc2mat on Unix command-line using the downloaded and formatted textfile of phenotypes:

```
./doc2mat -nostem -mystoplist=<downloaded list from PubMed> -skipnumeric -nlskip=1 -tokfile
<phenotypefile> <matrixfile>
```

A detailed help to all parameters can be found at:
http://glaros.dtc.umn.edu/gkhome/files/fs/sw/cluto/doc2mat.html

6. With the resulting matrixfile, use the CluTo package to cluster the phenotypes:

```
./vcluster -colmodel=<string> -crfun=<string> <matrixfile> <nclusters>
```

Our parameters:
colmodel=tfidf
crfun= I2
nclusters=1000

Two other very useful optional files can be created, namely the rlabel file (giving the phenodocs identifiers in order of appearance in the matrix file) and the clabel file, containing all unique words as represented by feature identifiers in the matrix file.

7. The CluTo result file contains the same number of lines as the input file. In each line of the result file, an integer indicates for the phenotype in the same line from the input file to which cluster (number) it belongs. With this, and using the rlable, the clable and the tokfile, the clustered phenotypes can be reconstructed from the CluTo result.

8. By their associations to the Entrez Gene identifiers, this clustering of phenotypes induces a highly coherent functional clustering on genes.