Exercise 3. Cluster and function enrichment analysis

1. Install the following software on your desktop computer. The two software are available for both Windows and Mac version.
   Cluster 3.0: http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm
   Java TreeView: http://sourceforge.net/projects/jtreeview/files/jtreeview/

2. Prepare the data files.
   - In this project, RNA-seq experiments were done on samples representing four different maize leaf development stages, from young to mature leaf tissues (Sample names are s1, s2, s3, and s4. *Nat Genet.* 2010, 42:1060-7). Using FileZilla to download the files from the BioHPC server, in the directory "/shared_data/RNAseq/exercise3/". There are two files in the directory: "genes.txt" and "maize.annot". The "genes.txt" file is a table of normalized expression levels.
   - You can examine the file "genes.txt" with Excel. When working with your own data, this file can be created with Excel.
   - Add +1 to all values in the excel table, so that there is no "0" in the table. This step is necessary as we will do log transformation later.
   - Save the Excel file as a "Text (tab delimited)" file.
3. Run Hierarchical Cluster.
   - Open the data file in Cluster.
   - Do Log transformation: click "Adjust Data" tab; check "Log transform data"; click "Apply".
   - Filter Data: click "Filter Data" tab; check the filter item 3 and 4; change "observations with abs(Val)>=" to 5; change "MaxVal-MinVal>=" to 2.0. Click "Apply filter", followed by "Accept Filter". As the filter is applied on log transformed data, this setting will keep the genes with FPKM value above 2^5 in at least one of the samples, and fold change above 2^2 between highest and lowest samples. The filtering is arbitrary, we would like to keep the filtered gene number below 10,000 after filtering.
   - Center genes: click the "Adjust Data" tab; uncheck the "Log transform data" box if it is still checked, as we do not want to do log transformation twice; check "Center genes" and "Median"; Do NOT check "Normalize genes" as the FPKM values are already normalized. Click "Apply".
   - Run Hierarchical clustering: check "Cluster" both under Genes and Arrays; click "Average linkage".
4. Run K-Means cluster
   - Click the "K-Means" tab.
   - Check "Organize genes". Change "number of clusters(k)" to 12. Click "Execute". There is a new ".kgg" file is created, you can open this file in Excel. This file has two columns: gene name and cluster ID. In this file, all the genes in your data set were separated into 12 clusters based on their expression patterns across the 4 samples. The "cluster ID" column indicates what cluster each gene is in.

5. Visualize the hierarchical clustering results.
    - Open the ".CDT" file in TreeView. The manual is available at http://jtreeview.sourceforge.net/manual.html.
6. Function enrichment analysis with BLAST2GO.
    - Get the Blast2Go Basic software. From the BLAST2GO web site: http://www.blast2go.com/, click "Blast2GO"->"Free Blast2GO Basic". You will need to fill out the registration form to get the software activation code. Then click Download button to download the right version of Blast2Go for your computer.
    - Start Blast2GO and load the maize Gene Ontology (GO) annotation file. In Blast2GO menu, click "File"->"Load"->"Load Annotations", and load the maize.annot file you downloaded at step 2. This file was created with Ensembl BioMart web tool (www.ensemble.org for animal species or plants.ensembl.org for plant sepecies).
    - Using Excel to open the .kgg file you created at step 4 (K-means clustering). Prepare two new excel files with just lists of gene names: 1)a file with all gene names (first column only) in the .kgg file; 2) a file with only gene names with second column value=0. Save both files as tab-delimited text file in Excel.
    - After you finishing load the maize.annot file into BLAST2GO, click "Analysis"->"Enrichment Analysis". You will need to provide the files with reference gene list (all genes you created in the previous step), and test gene list (genes in group 0), and click run. The results is a table with GO functions over represented in group 0. You can save the results as a text file, which can be opened in Excel.