CS 4370/5370 Project P4 Instructions

Due at the start of class Tuesday, Nov. 29th. Late submissions are allowed **ONLY** during the three days following the due date, with 5%, 10%, and 15% penalty, respectively.

Objective: Order data columns, remove genes and individuals with excessive missing data, and discretize data.

Install software:

1. Log onto '`tc.rnet.missouri.edu`' using your SSO credentials
2. Download the following and move to appropriate directories
   a. `extractCols_v1.0.tar.gz`
   b. `miss_v1.0.tar.gz`
   c. `discretize.tar.gz`
3. Unzip: '`gunzip` *`filename`*`.tar.gz`'
4. Untar: '`tar -xvf` *`filename`*`.tar`'
5. Move to source code: '`cd` *`filename`*'
6. Compile: '`make`'

Download data, fill empty cells and blanks, copy data table into text editor: *(You have already done these steps!)*

1. Download data from GEO Omnibus (https://www.ncbi.nlm.nih.gov/geo/)
   a. Enter your GEO accession number in search box
   b. Open a document to record information about your data for future reference
      i. Copy and paste the following in your document:
         1. GEO accession number
         2. Title
         3. Organism
         4. Experiment type
         5. Summary
         6. Overall design
         7. Contributors
         8. Citation (download paper if available)
         9. Submission date
         10. Last update date
         11. Contact name
         12. E-mails
         13. Phone
         14. Organization name
         15. Platforms
   c. Scroll down to "Download family" section and click on "Series Matrix Files"
   d. Move downloaded file to an appropriate directory
2. Make a copy of the datafile to keep as backup
3. Unzip the data file
   a. Type 'gunzip *filename*'
4. Open data in spreadsheet and remove comment rows
   a. Comment rows start with '!'
      i. Many comment rows before the data table starts
      ii. Note there is a comment after the end of the data table – be sure to remove it!

5. Replace empty cells with NA
    a. 'Select All' to highlight the data table
    b. 'Replace All' with nothing (don't type anything, not even a space) in the find box and 'NA' in the replace box
    c. Note the number of values replaced in your information record
6. Replace spaces in header rows and columns with underscores
    a. Highlight header rows
    b. 'Replace All' with a single space in the find box and a '_' in the replace box
    c. Repeat for header columns
7. Copy data into plain vanilla text file
    a. 'Select All' to highlight all data
    b. Copy and paste into a plain vanilla text editor (e.g. mobaxterm editor, emacs, vim - do not use Notepad)
    c. Save with an appropriate name (e.g. short trait word)
    d. Record the numbers of data rows and columns and header rows and columns in your information record

Obtain lists of Cases and Controls IDs: *(You have already done this step!)*

1. Comment rows of your data file include phenotype status
2. Determine the correct labels for your classification of 'Cases' and 'Controls'
    - May need to refer to manuscript
3. Copy the relevant row in your data
4. Copy the header row from your data table
5. Paste both in a separate spreadsheet
6. Transpose to columns
7. Sort based on phenotype column
8. Carefully check that 'Cases' are separated from 'Controls'

Reorder columns in data table:

1. Create two new empty files by typing: '`touch casesIDs.list`' and '`touch ctrlsIDs.list`'
2. Copy Cases IDs and paste into '`casesIDs.list`'
3. Copy Controls IDs and paste into '`ctrlsIDs.list`'
4. Use `extractCols` to extract data for cases into '`case.txt`'.
5. Change `WRITEHEADCOLS` to 0 in '`extractCols.h`' and recompile
6. Use `extractCols` to extract data for controls into '`ctrl.txt`'.
7. Paste the two files together: '`paste case.txt ctrl.txt >> yourSortedFilename.txt`'
8. Use word count '`wc filename`' to check that the sorted file has the same number of rows as the original
9. If you didn't drop some of the individuals from your case/control lists, the number of strings should be the same

Remove excessive missing data:

1. Install '`miss`' program
    a. Test on small file
        i. Look at small file: 'cat small_gex.txt'
        ii. Count number of rows and columns for headers and data
        iii. Run program: './miss small_gex.txt 10 10 1 1'
        iv. Follow prompts to remove ***either*** rows or columns by selecting values less than 100
            1. Use value of 100 for the dimension not being reduced (row or column)

v. Iterate until no more than 5% missing for all rows and no more than 5% missing for all columns

2. Run program on your data
    b. Try to preserve as many individuals (columns) as feasible
        i. To start, only remove the individuals with extremely high missing rates
        ii. You might try running from the start a couple of times using different scenarios

Discretize your data:

1. Install '`discretize`' program
2. Run program on your data using 0.3 for the `percent` value
3. Make a screen shot of the program output

Submit on Canvas:

1. Screen shot of the word and string counts for your data file before sorting (but after comment rows removed)
2. Screen shot of the word and string counts for your data file after sorting
3. Log file that was automatically generated when you ran the '`miss`' program
4. Screen shot of the output for the '`discretize`' program