Implementation of SAM (Significance Analysis of Microarrays) in BioBIKE

I. Preliminaries: Read in and set up data

A. DEFINE tusher-data as the contents of a file called "tusher-data.txt", the data used by Tusher et al in their article

Use the READ function with SHARED, TABBED, and CONVERT-NUMBERS options
- The SHARED option reads the file from the shared-file directory, accessible by everyone.
- The TABBED option reads it in as a tab-delimited table
- The CONVERT-NUMBERS option converts all strings that look like numbers into numbers

Note in the result window that the data comes as a header followed by each line of data, one line for each gene.

```
(DEFINE tusher-data =
  (READ FROM "tusher-data.txt" SHARED TABBED CONVERT-NUMBERS))
```

B. DEFINE the headers as the first line of tusher-data, from column 2 to the end

Use the INSIDE-LIST function to get columns 2 to the end.
Use the FIRST function to get the first line

```
(INSIDE-LIST (FIRST tusher-data) FROM 2)
```

C. Re DEFINE tusher-data to be all lines except for the header line

Use the INSIDE-LIST function to get lines 2 to the end

```
(DEFINE tusher-data = (INSIDE-LIST tusher-data FROM 2))
```

D. DEFINE tusher-groups as the 36 groups of indices used by Tusher et al as the contents of the file called "tusher-groups.txt"

Use the READ function with SHARED, TABBED, and CONVERT-NUMBERS options

II. Extract data for one of the 36 groups of indices

A. DEFINE group-indices as the FIRST set of indices in tusher-groups, consisting of the indices of the first group (1 2 5 6) and the indices of the second set (3 4 7 8)

```
(DEFINE group-indices = (FIRST tusher-groups))
```

B. DEFINE group1-indices as the FIRST set of indices from group-indices (1 2 5 6)

```
(DEFINE group1-indices = (FIRST group-indices))
```

C. DEFINE group2-indices as the SECOND set of indices from group-indices (2 3 7 8)

D. DEFINE gene-data as the FIRST set of data within tusher-data, consisting of the gene name and eight expression values.

E. DEFINE expression-values as the items in gene-data from the second one to the end, using INSIDE-LIST as before.

F. DEFINE group1 as the items in expression-values corresponding to the group1-indices

(DEFINE group1 = (INSIDE-LIST expression-values ITEM group1-indices))

G. DEFINE group2 as the items in expression-values corresponding to the group2-indices

III. Package Section II into a function that takes two sets of indices and a list of expression values and returns two sets of expression values

A. DEFINE-FUNCTION Expression-values-of by bringing down DEFINE-FUNCTION and providing Expression-values as its name.

B. Provide a SUMMARY by clicking the SUMMARY arrow, clicking SUMMARY and then describing the function any way that makes sense to you.

C. Provide the name of the first required argument, call it group-indices

D. Add another required argument, call it gene-data.

E. Cut and paste each command from **II.B, II.C., and II.E. through II.G** into forms in the BODY. Add more forms as needed.

F. Add one more form, and put in it:

(LIST group1 group2)

G. Execute the function. You should now have EXPRESSION-VALUES-OF in your FUNCTION button.

H. Test the function by bringing it down and giving it sets of indices and a set of expression values. Try:

(EXPRESSION-VALUES-OF ((1 2 5 6) (3 4 7 8)) gene-data)

IV. Calculate statistical quantities based on the two groups

A. DEFINE mean1 as the MEAN of group1 expression values.

(DEFINE mean1 = (MEAN group1))

B. DEFINE mean2 as the MEAN of group2 expression values

C. DEFINE s0 as the constant 3.3, empirically found by the authors to minimize the variability of s (see below).

D. RUN-FILE "gene-specific-scatter.bike" to bring in Tusher et al's s(*i*) function into your space.

(RUN-FILE "gene-specific-scatter.bike" SHARED)

Note that you have gained a new button called FUNCTIONS that contains this function.

E. DEFINE d as:

(DEFINE d = (DIVIDE (SUBTRACT mean1 BY mean2)
BY (ADD (GENE-SPECIFIC-SCATTER group1 group2) TO s0)))

V. Package Sections IV into a function that calculates d as a function of two groups of expression values

A. DEFINE-FUNCTION d-of as before.

B. Provide a SUMMARY as before

C. Provide the name of the required argument, call it groups

D. In the form in the BODY section, define group1

(DEFINE group1 = (FIRST groups)

E. Add another form in the BODY section, define group2 as the SECOND of the groups

F. Cut and paste each command from **IV.A through IV.E** into forms in the BODY. Add more forms as needed.

G. Add one more form, and put in it: d

H. Execute the function. You should now have D-OF in your FUNCTION button.

I. Test the function

VI. Calculate the expected relative difference $d_E(i)$ for a given gene

$d_E(i)$ is the average $d(i)$ calculated for all 36 sets of indices

A. DEFINE d-sum as the sum of all $d(i)$ values over all sets of indices. Do this by applying the D-OF function to all sets of indices:

```
(DEFINE d-sum
   =  (APPLY-FUNCTION-OF indices
         = (D-OF (EXPRESSION-VALUES-OF indices gene-data))
         TO tusher-groups))
```

B. Calculate the average by dividing d-sum by 36

C. Package the calculation of the average into a function, called Expected-rel-diff, giving as required argument gene-data

VII. Compare the expected relative difference to the actual relative difference

A. Go through the first 10 genes, either using APPLY-FUNCTION-OF or FOR-EACH, calculate the relative difference (through D-OF) and the expected relative difference (through EXPECTED-REL-DIFF). Collecting these quantities for all the genes and plotting them gives you Fig. 3A of the article.

B. If the distance between the relative difference and expected-relative-difference is greater than a threshold (the article gives 1.2 as an example), then get a count of those genes "called significant".