Program Descriptions

This document describes the programs that were used in the publication *HIV Infected CD4+ T Cell Clones Are More Stable than Uninfected Clones During Long-Term Antiretroviral Therapy*. Shuang Guo, Brian Luke, Amy R. Henry, Samuel Darko, Leah D. Brandt, Ling Su, David Sun, Daria Wells, Kevin W. Joseph, Dimiter Demirov, Elias K. Halvas, Daniel C. Douek, Xiaolin Wu, John W. Mellors, Stephen H. Hughes. **PLoS Pathogens**, 2022.

Each program has been compiled using GNU Fortran, but it should work with any compiler that contains the GETARG routine.

All programs were written by Brian T. Luke, Ph.D. and all enquires should be addressed to [Brian.Luke@nih.gov](mailto:Brian.Luke@nih.gov).

**weightfit4**

*Summary*

This program calculates the difference in clone sizes between two time points. Three difference metrics are explored, and this program was used at the initial stage of our study to compare the difference metrics and examine the effects of using different numbers of clones in the comparison. This program was not used in the publication sited above but represents a starting point where we explored the effect of using different numbers of the largest clones to determine the effective difference, as well as explore different difference metrics.

*Running the Program*

{Path to executable}weightfit4 INFILE OUTFILE NPICK

For example,

weightfit4 example\_IS\_clone.txt example\_IS\_weightfit4.txt 20

*Details*

INFILE is the input file in tab-delimited format that contains a minimum of three columns. The first column is the label for the insertion site (IS) or T cell receptor (TCR) clone, and the next two columns contain the clone sizes at two time points. The first line of this input file contains the column labels. In the example input file, example\_IS\_clone.txt, the first line is

Clone\_Label Examp\_2014 Examp\_2015 Examp\_2017

This file contains four columns, but weightfit4 will only compare the clone sizes between 2014 and 2015.

OUTFILE is the name of the file that contains the comparison across the two time points.

NPICK is the number of largest clones to select from each time point. All clones with a size equal to the smallest selected clone are also included, and then the union of these sets is used.

In the example given above, the 20 largest clones (NPICK) are selected from the 2014 time point, and then all other clones that are as large as the 20th are also included. This results in a total of 34 clones being selected. The same procedure is used for the 2015 clones and a total of 21 are selected. These two lists are then merged, resulting in 43 unique clones. This means that 12 of the 21 largest 2015 clones are also among the 34 largest 2014 clones.

To mitigate differences in the total number of insertion sites or T cell receptors between the two time points, the clone sizes at each time point are scaled so that they both have an average clone size of 20. This value of 20 is completely arbitrary but is fixed in the program so that the calculated difference between one pair of time points can be compared to another pair of time points. Also note that this allows the input to be either observed clone sizes, where the number of insertion sites or TCRs can vary greatly from one time point to the next, or clones that are pre-scaled to a constant total size. As such, the clone sizes are integer values on input, but are converted to floating-point numbers when they are scaled by this program.

These 43 scaled clones are effectively placed on a scatter plot and five different regression lines are calculated. The first is an unweighted regression line, while the other four are weighted regression lines. The weighting of each clone is given by the mean size of the clone for the two time points raised to a given power. Therefore, the five regression lines can be considered as weighted by the mean clone size raised to the power of 0.0 (unweighted), 0.5, 1.0, 1.5 and 2.0. As the exponent increases, more weight will be given to the larger clones relative to the smaller ones.

The three difference metrics are based on the perpendicular distance between the point on the effective scatter plot and a given regression line. The first is the root mean squared error (RMSE), which is simply the square-root of the mean distance-squared to the regression line across the selected clones. The second difference metric is the average absolute error (AAE) which is the mean perpendicular distance between each point on the scatter plot and the given regression line. The final difference metric is simply called the relative error (RELE) and is the distance-squared to the regression line divided by the mean of the scaled size of this clone at the two time points, averaged over all selected clones.

For each weighting of the regression line, the program prints out the intercept and slope of the regression line as well as the RMSE, AAE and RELE. The program then prints out the scaled sizes of each clone at the time points as well as the clone sizes at the second time point as predicted by the five regression lines. Finally, the selected clones as well as their initial sizes are listed. The example output shows that one of the large clones selected at the first time point is not observed at the second, while two of the large clones at the second time point are not observed at the first.

**IScomp3**

*Summary*

This program reads an NCOL+1 column tab-delimited file that contains the 100 largest HIV insertion site clone sizes at NCOL time points and the time-average clone size. It runs 10,000 trials where 2,000 insertion sites are probabilistically selected from these clones and the difference between the sets of clone sizes is determined. This is run NCOL\*(NCOL+1)/2 times so that each time point is compared to itself and to all other time points.

*Running the Program*

{path to executable}IScomp3 INFILE PREFIX NCOL SEED

For example,

IScomp3 example\_IS\_100largest.txt example\_IScomp3\_ 3 511900173.0

*Details*

INFILE is the input file. The input file example\_IS\_100largest.txt is a tab-delimited file with four columns. The first three columns contain the size of each of the 100 largest clones at each time point and the fourth column contains the average size over the three time points. This fourth column was used to sort the clone sizes from largest to smallest. This file only contains the 100 largest time-averaged clones. It could contain more than 100 clones, but only the first 100 will be used.

PREFIX is the prefix for the output files.

NCOL is the number of time points being considered; meaning that INFILE contains NCOL+1 tab-delimited columns.

SEED is the seed to the included random number generator and should be large and odd, preferably a prime number.

After reading in the clone sizes for the 100 largest time-averaged clones at the NCOL time points, the program builds a roulette wheel for each time point. Each wheel contains 100 wedges, and the size of each wedge is equal to the fractional size of that clone at that time point. Therefore, the sum of all wedge sizes is 1.0 and a random number between 0.0 and 1.0 determines a particular wedge, or clone.

If a time point is compared to itself, the same roulette wheel is used to independently select two sets of 2,000 insertion sites from the 100 clones. If two time points are compared, the roulette wheels for these times are independently used to select the 2,000 insertion sites. The sizes of each of the 100 clones are compared on a scatter plot and are fitted with an unweighted or weighted regression line. The perpendicular distance from each point is used to calculate the relative error (RELE) as described more fully in the details of *weightfit4*. This is repeated 10,000 times, so the output contains a section with five columns representing the weighting of the regression line, and 10,000 rows giving the RELE for each regression line.

Since there are three time points in the example data, there are a total of 3\*4/2 or six within-time and between-time comparisons. The PREFIX is used to generate the name of the output file, and in this example the names will be example\_IScomp3\_A.txt through example\_IScomp3\_F.txt. To compare the first time point against itself versus the first against the second time point, copy Column A, Rows 6 through 10,005 from example\_IScomp3\_A.txt and example\_IScomp3\_B.txt into a separate file and run a 2-sample t-test using any standard statistical package. Each of these columns can also be used to generate a histogram or kernel density plot to show the variation in the calculated differences.

**TCRcomp3**

*Summary*

This program is quite similar to IScomp3, in that it determines the difference in clone sizes within and between time points in 10,000 trials. The main difference is that there are thousands of TCR clones that span the size of the 100 largest IS clones. Therefore, in each trial a master roulette wheel is used to select the 100 clones, and then these clones are used to build roulette wheels for each time point being considered.

*Running the Program*

{path to executable}TCRcomp3 INFILE PREFIX NCOL SEED

For example,

IScomp3 example\_TCR\_100matched.txt example\_TCRcomp3\_ 3 284900173.0

*Details*

INFILE is the input file. The input file example\_TCR\_100matched.txt is a tab-delimited file with four columns. The first three columns contain the size of each of the clones at each time point and the fourth column contains the average size over the three time points. This fourth column was used to match the time-averaged TCR clones with the 100 largest IS clones.

PREFIX is the prefix for the output files.

NCOL is the number of time points being considered; meaning that INFILE contains NCOL+1 tab-delimited columns.

SEED is the seed to the included random number generator and should be large and odd, preferably a prime number.

For the IS clones (see *IScomp3*) we took two sets of 2,000 insertion sites from the 100 largest clones and determined the difference between them. We want to do the same for comparably sized TCR clones, but there are thousands of TCR clones that span the range of these IS clones. Therefore, this is done as a two-step process. the first step is to use the time-averaged clone sizes to create a master roulette wheel for that dataset. For each comparison, within a time point and between time points, this master roulette wheel is used to select 100 TCR clones. For each time point this set of 100 clones is used to construct a secondary roulette wheel, if it is a comparison within a single time point, or two secondary roulette wheels for the two time points being compared. For a comparison within a time point, the single secondary roulette wheel is used to independently generate two sets of 2000 TCRs, while for a comparison between time points, a secondary roulette wheel is constructed for each time point and they are both used to construct independent sets of 2000 TCRs. Each set is then used to determine the clone sizes, the 100 clones are placed on a scatter plot and the one unweighted and four weighted regression lines a fit to the 100 points (see description of *weightfit4*). The output for each comparison consists of the RELE difference between the sets for each of the regression lines for each of the 10,000 trials.

Since there are three time points in the example data, there are a total of 3\*4/2 or six within-time and between-time comparisons. The PREFIX is used to generate the name of the output file, and in this example the names will be example\_IScomp3\_A.txt through example\_TCRcomp3\_F.txt. To compare the first time point against itself versus the first against the second time point, copy Column A, Rows 6 through 10,005 from example\_TCRcomp3\_A.txt and example\_TCRcomp3\_B.txt into a separate file and run a 2-sample t-test using any standard statistical package. Each of these columns can also be used to generate a histogram or kernel density plot to show the variation in the calculated differences.

**sizetest**

*Summary*

This program is used to determine the effect of a particular sample size (e.g., number of observed genomic HIV insertion sites) on observing the largest clones. It reads in a base distribution of relative clone sizes and probabilistically selects a user-supplied number of samples. The output consists of a table showing the number of times each of the 100 largest clones was selected in each of the 10,000 trials.

*Running the Program*

{path to program}sizetest INFILE OUTFILE SEED NSAMP

For example

sizetest example\_total\_obs.txt example\_sizetest\_500.txt 250042713.0 500

*Details*

INFILE is a single-column file that contains the relative size of each clone/single. In the included example input, there are 12,032 unique insertion sites (clones/singles) representing a total of 19,717 “observed” insertion sites.

OUTFILE is a tab-delimited file containing the number of times a member of the 500 largest clones is selected in each of the 10,000 trials.

SEED is the seed to the random number generator. It should be large and odd, preferably a prime number.

NSAMP is the number of samples (observed insertion sites) taken from this overall distribution.

The output (e.g., example\_sizetest\_500.txt) contains three lines that list the input file, number of samples selected in each trial, and the seed to the random number generator. This is followed by a 10000x100 table (with column headers) that lists the number of times each of the 100 largest clones was selected in each of the 10,000 trials.

Subsets of this table can be used to determine, for example, the number of times the 2nd through 10th largest clone is seen at least as many times as the largest clone, and the number of times each of the 10 largest clones is missed.

**missing2**

*Summary*

This program reads the output of *sizetest* and determines the number of times each of the N largest clones is not seen in each of the 10,000 trials; N = 10 to 100 in steps of 10.

*Running the Program*

{path to executable}missing2 INFILE OUTFILE

For example

missing2 example\_sizetest\_500.txt example\_missing2\_500.txt

*Details*

INFILE is the name of the output file from sizetest (e.g., example\_sizetest\_500.txt).

OUTFILE contains the total number of insertion sites seen in the N largest clones and the number of these clones that are not seen.

Using the example above, the first trial gives the following results for up to the top 40 clones.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| IS\_10 | MISS\_10 | IS\_20 | MISS\_20 | IS\_30 | MISS\_30 | IS\_40 | MISS\_40 |
| 35 | 0 | 48 | 2 | 51 | 9 | 59 | 15 |

This states that in the first trial of *sizetest*, 35 of the 500 insertion sites were in the 10 largest clones, 48 in the 20 largest clones, and so on. In this first trial, none of the 10 largest clones were missed, 2 of the 20 largest clones, 9 of the 30 largest, and 15 of the 40 largest clones were missed by this set of 500 insertion sites.

There are a total of 20 columns to account for up to the 100 largest clones, and there are 10,000 rows: one for each *sizetest* trial.