(21 Apr 2017)

MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics
# Projects: Statistical Assessment of Metagenome Populations (uses ViroBIKE)

It's pretty common to find yourself with a compelling need to compare two populations in some respect or another, making use of this or that statistical test. It's easy to mix them up and even easier to have no convincing way to tell if you did. But… if you can program a computer, then you often have the option of throwing all those tests against the wall and doing a direct test: If I randomized the data, how often would I get the result I actually got?

Here we do just that, using two virus populations taken from the South Sea.

A. **Simulation of a statistical test**

Suppose that you are interested in the diversity of viruses in the world's coral reefs. You intend to sample viruses in various coral reefs and use their sequences to assess viral populations. You have chosen coral reefs adjoining two nearby south pacific islands (see figure at right), part of the Line Islands. One, Tabuaeran, is part of the Republic of Kiribati. The other, Palmyra, is a U.S. possession. Tabuaeran has a population of about 2500, while Palmyra is nearly uninhabited. It would be of interest to learn whether the coral reefs and the viruses that infect them differ in a nearly untouched reef as compared to one in which there has been significant human activity.[*]
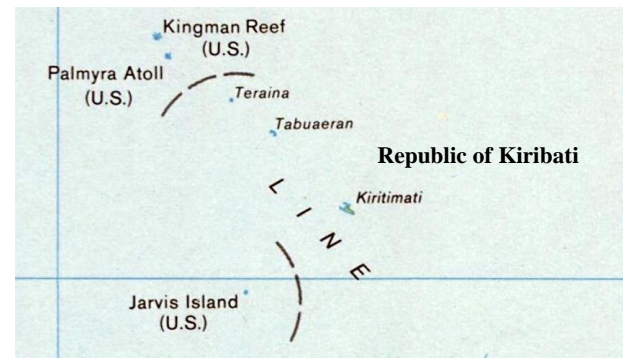
Your plan is to compare viral sequences from each site. It would take about $40,000 to get good sequence coverage of a metagenomic sample, and you don't have that kind of money. Instead, you'll sequence clones of 200 viral sequences from each site and hope that fairly represents the total viral population.

A1. Before embarking on this project, let's step back and take a God-like view of the populations of viral sequences from the two sites. We'll do this by displaying the GC-fractions of metagenomes[†] from Palmyra and Tabuaeran (nicknamed Tabu).

  - Go to ViroBIKE (from the usual BioBIKE portal, click the ViroBIKE public site).

  - Bring down the PLOT function (either from the Input-Output button or from the alphabetical list).

  - This function requires a *data-list*, which can either be a list of values (to get one line on the graph) or a list of lists of values (to get multiple lines). We want to plot two sets of GC-fractions simultaneously (from Palmyra and Tabu), so select the *data-list* argument box and bring down the LIST function.

---

[*] Those hungry for more about coral reefs and microbial diversity can check out Dinsdale EA et al (2008). Microbial Ecology of Four Coral Atolls in the Northern Line Islands. PloS ONE 3(2):e1584.

[†] A metagenome is a collection of sequences obtained from a sample that contains multiple – perhaps thousands – of different organisms. It is not generally possible to piece together from the reads the complete genomes of the organisms in the sample, so what you get are individual reads or usually short contigs.

- Add another item so that LIST has two *item* boxes, one for each island.

- Bring down GC-FRACTION-OF into the first *item* box.

- In the *entity* argument box of GC-FRACTION-OF, bring down CONTIGS-OF (from either the Genome button or the alphabetical list). A contig is a fragment of DNA. All the CONTIGS-OF `Palmyra` would give you all the sequenced DNA fragments from phages isolated from Palymyra. How many contigs are there from Palymyra? (get a COUNT-OF them)

- In the *entity* box of CONTIGS-OF type `Palmyra` (<u>not</u> in quotes). Execute the function (should take ~10 seconds to run). You are hoping to get a GC-FRACTION-OF each contig, so lots of numbers. Is that what you get? If you get one number, something's wrong. That number is the mean GC-fraction of all contigs taken together. If you want the individual GC-fractions (and you do), use the EACH option.

- Copy the GC-FRACTION-OF function with everything inside it and paste it into the other *item* box of the LIST function. Change `Palmyra` to `Tabu`.

- If you execute the function, you'll get a smush of many tens of thousands of points (or you'll time out trying). We need to organize the points by putting them into bins, counting how many contigs have GC-FRACTIONS between .30 and .31, how many between .31 and .32, etc. Mouse over the Options Icon of the PLOT function and select BIN-INTERVAL. Enter 0.01 into the *value* box.

- And for sake of esthetics, go back to the Options Icon and select X-LABEL. Type "GC Fraction" in the *value* box. Select Y-Label, and type "Frequency" in the *value* box.

- <u>Now</u> execute the function.

Does the solid line (Palmyra GC-fractions) coincide with the dashed line (Tabu GC-fractions? Which one is shifted to the right, i.e. to higher GC-fractions? What is the difference in the means of these two populations? 1.6% doesn't sound like a huge difference, but when you're looking at so many contigs (COUNT them if you don't believe me), it may well be significant.

A2. Back to real life. You don't have that picture of the viral sequence populations at the two sites – not enough money to get the sequences. Instead you're sequencing 200 viral fragments from each site. Let's simulate that situation by getting 200 samples each.

- DEFINE `palmyra-seqs` as the FIRST 200 CONTIGS-OF `palmyra`.

- Likewise, DEFINE `tabu-seqs` as the FIRST 200 CONTIGS-OF `tabu`.

- Redo the PLOT from **2a**, changing `palmyra` to `palmyra-seqs` and `tabu` to `tabu-seqs`. (Actually now that you have the sequences, CONTIGS-OF is no longer necessary. But it doesn't hurt).

- Oooh, spiky! Change the BIN-INTERVAL to .02 and try again.

- That's better, but is there any difference between the two? Calculate the difference in the mean GC-FRACTIONs as in **2a**, but with `palmyra-seqs` and `tabu-seqs`. ***Remember this number!***

There's still a difference, but it sure doesn't look as good on the graph!

A3. How can we test whether the difference between the two means is significant in this sample (as, in our god-like state, we know it is in the full population)? Suppose there is no real difference. Suppose that the two islands have the same viruses, and when you sample at Palmyra and at Tabu, you're really sampling from the same pool. Let's simulate that situation.

- DEFINE `single-pool` as both `palmyra-seqs` and `tabu-seqs` JOINed together. Notice that the result (in the RESULT window) shows the Palmyra sequences in order. The Tabu sequences follow them (off screen).

- DEFINE `randomized-pool` as `single-pool` RANDOMIZEd. You'll find the RANDOMIZE function under the List-Tables, List-Production. Notice that the result shows seemingly random sequences. It is actually `single-pool` in a random order.

- DEFINE `set1` as the FIRST 200 sequences of `randomized-pool`.

- DEFINE `set2` as the LAST 200 sequences of `randomized-pool`.

If Palmyra and Tabu have the same pool of viruses, then taking the first 200 of a mixed collection is no different from taking the last 200.

- DEFINE `mean1` as the MEAN of the GC-FRACTIONs-OF `set1`.

- DEFINE `mean2` in an analogous fashion.

- DEFINE `difference` as one mean SUBTRACTed from the other. Actually (since you just want to know if there is a significant difference between the Palmyra and Tabu means regardless of direction) surround the SUBTRACT function by ABS (absolute value).

Is the difference of the means GC-FRACTIONs of these two random subsets smaller than what you got with the real `palmyra-seqs` and `tabu-seqs`? Maybe yes, maybe no. What we really need to know is how <u>frequently</u> would random subsets of a single pool have a difference of means greater than what you observed with `palmyra-seqs` and `tabu-seqs`.

A4. Here's how we'll determine whether the observed difference in means is greater than what you'd expect by chance.

- Set up a PROGRAM (brought down from the Define button), consisting of the operations of **2c**. Excluding the definition of `single-pool`, there were six DEFINE operations. Use the More Icon to give PROGRAM six *form* boxes.

- Drag each of the operations of **2c** (except the definition of `single-pool`), in order, into the appropriate *form* box of PROGRAM.

- Execute the PROGRAM. It should give you the difference of means of two random subsets of `single-pool`, different from the difference you got in **2c**. Why is it different?

- Collapse the program (from the Action Icon at the upper left of PROGRAM, giving it a name that makes sense to you, maybe something like difference between sets.

2e. Now that you know how to generate random subsets and determine the difference in their mean GC-fractions, you'll do it 100 times and ask how many of those times is the

difference greater than the mean difference you observed with the Palmyra and Tabu samples.

- Bring down a FOR-EACH loop.
- This loop will have a Primary Control section, a Body section, and a Results section. Hide the other four sections.
- Open up the Body, and drag the PROGRAM box into the *body* form.
- From the Primary Control menu, choose *number* FROM *n1* TO *n2*.
- Add values to the *var*, *first value*, and *last value* boxes so that the variable `trial` goes from 1 to 100.
- You want to count the trial only if the difference between mean1 and mean2 is less than the difference of the means you actually observed in **2b**. Choose from the Results menu *when... count*.
- For *condition*, choose TRUE? from the Flow-Logic button. Click the *TEST?* icon and select >. Put `difference` (the variable you defined in your program) in the *value1* box of TRUE?, and put the difference of means you calculated in **2b** in the *value2* box. Be sure the difference is a positive number.
- In the *value* box governed by COUNT, put `trial`. When `difference` is greater than the observed difference, you'll COUNT the trial.
- Execute the FOR-EACH loop. How many times out of 100 trials was the difference in means of the random subsets greater than the observed mean?
- Execute the loop again. Now how many times?
- (brought down from the Define button), consisting of the operations of **2c**. Excluding the definition of `single-pool`, there were six DEFINE operations. Use the More Icon to give PROGRAM six *form* boxes.
- What conclusion can you draw from this simulation?

B. <u>Statistical shortcut</u>

That (Problem A) was a lot of work Surely there's an easier way to assess whether two sets of numbers are significantly different or instead are consistent with coming from a common pool. Is there an easier way to do the test of significance you performed in **Section A**? Yes indeed! Actually, it's essentially the same thing you did but with a major shortcut.

B1. Try the following:

- Calculate the MEANs and STD-DEVs (standard deviation) of `palmyra-seqs` and `tabu-seqs.`
- Go to the t-test calculator (see the Links and Resources page of the course web site).
- Check the "Enter mean, SD, and N" radio button.
- Enter the mean, SD (standard deviation), and N (number of sequences) for group 1 (`palmyra-seqs`) and group 2 (`tabu-seqs`). Click `Calculate now.`
- Find the **P value**. How does it compare with the number of times the difference between the GC fractions of the random subsets exceeded the observed difference comparing `palmyra-seqs` and `tabu-seqs`?

- Note the t score and the degrees of freedom (df). The latter is always the sum of the N values minus 2.[*]
- Go to the t-test probability calculator, also available from the Links and Resources page.
- Type in the t-value and the degrees of freedom. Then click Calculate. You should get the same P value. A two-tailed P value of 30% means that about 30% of the time the difference between random subsets taken from the same population exceeds the observed difference.

B2. What is a t-test? Suppose a t-test gives a P value of 0.05. How would you put this value into a meaningful English sentence, using your experience from Section A.

C. **Simulation of another statistical test**

Sometimes a simulation is easier than a statistical test, but even when it's not, constructing a simulation often gives you an intuitive understanding of the test that may otherwise be lacking. If you gained some insight by simulating a t-test, you might want to gain similar insight into chi-squared tests. If so, try Was Mendel Right? (This one is much less work than Section A)

---

[*] Degrees of freedom… how many of the values in a sample can you make up without restriction, given that you know the mean of the sample? Suppose that there are 5 numbers in a sample and the mean is 10. The first number can be anything, say 47. The second number can be anything, say 29, but when you get to the 5th number, it **cannot** be anything, because the sum of the numbers has to add up to 5 times the mean of 10. So you have no choice for the 5th number. The number of values you can make up is $5 - 1 = 4$. In general, the number is $n - 1$ for each sample. If you have two samples, then it's $N_1 - 1 + N_2 - 1 = N_1 + N_2 - 2$.