Editor's Corner

Mutation Rates – Who's Got the Right Values?

A discussion on Y-STR mutation rates seems to come up on the various e-mail lists about once each month. The discussion has also been carried on from time to time in the professional journals as well.

Most of the interest has been in the rates derived from father-son pairs, as that seems most applicable to Y-chromosome surname projects. Shortly after Family Tree DNA (FTDNA) expanded its offerings to 37 markers about three years ago, the company asked selected surname project administrators to submit mutation data and genealogies from their projects where the genealogy was fairly well established. They have never published their results, so we don't know how large of a dataset they used, but they have announced the average values that they obtained. Since FTDNA used three panels of markers at that time, there were five possible average values of interest—the average for markers 1-12, markers 13-25, markers 26-37, markers 1-25, and markers 1-37, and the values that were obtained, respectively, were .0039, .0048, .0075, .0044, and .0058 mutations per locus per generation or transmission. These values were a little surprising—most people were expecting somewhat lower values. A value of .0025 had been widely assumed for markers 1-25 prior to the FTDNA study.

To address the issue of mutation rates in an independent study, Charles Kerchner started his "mutation log" in 2005 where data from surname projects can be deposited. The overall goals and methods of the project are essentially identical to those of the FTDNA study, but this one is in the public domain where the numbers behind the averages may be seen. In order to submit data to the mutation log, the genealogy of the participants must be known to the project administrator and he must have reconstructed the ancestral haplotype so that mutations from that haplotype can be accurately counted. Mutations are only counted once when the same mutation is inherited by more than one participant. Sometimes the genealogy is not sufficiently well-known to make it clear which mutations were inherited from an ancestor and which have occurred independently, but Kerchner asks that in case of uncertainty, the data be left out.

Kerchner's study has been somewhat successful, but it is likely that there are many more projects with useful data existing in the community of surname projects than have been submitted. It is rather unfortunate that not every administrator has taken the trouble to submit his or her data, because the results could provide a very important check on the FTDNA study.

At the time of this writing, there have been 45 submissions from various surname projects. There are differing numbers of transmissions and mutations for each panel, but for example, overall on markers 1-37, there have been 75258 marker transmissions reported and 309 mutations have been observed, for an average mutation rate per marker on the 37-marker panel of 0.0041 ± .0002 (one standard deviation). The corresponding mutation rates calculated from similar data for the panels 1-12, 13-25, 26-37, and 1-25 are 0.0024, 0.0029, 0.0071, and 0.0027. Data on FTDNA markers 38-67 are just starting to be submitted, but it is obvious already that this panel has an average rate that is probably the lowest of the four panels.

There are some significant differences in the average mutation rates from the FTDNA study and the Kerchner study. FTDNA's rates are 40-60% higher—they are not within the error bars of the Kerchner rates (FTDNA hasn't divulged their error bars). There can be selection bias when the data are voluntarily self-reported, rather than being collected according to a predetermined sampling procedure. However, this problem apparently applies to both the FTDNA and Kerchner studies, though each may be affected in a different way.

Another approach to calculating mutation rates was published in the Fall 2006 issue of this journal by John Chandler (2006). Chandler's approach compares thousands of haplotypes and basically yields relative rates for the 37 FTDNA markers individually. These relative rates (all 37) can then be calibrated to the absolute rates that have been published for several of the markers, based on father-son pairs. Alternatively, these relative rates can be calibrated to any other absolute rates, such as the "effective rates" of Zhivotovsky, discussed below.

The average mutation rates for the 1-12, 1-25, and 1-37 panels were found by Chandler (for the father-son calibration) to be 0.00187±0.00028, 0.00278±0.00042, and 0.00492±0.00074. In this case the 95% confidence intervals for the Kerchner rates and the Chandler rates overlap, so to this point in the Kerchner study at least, the two studies are providing consistent results. This is quite important since the two approaches to calculating the rates are totally different and independent. However, Chandler's rates are even further from those

of FTDNA than are Kerchner's. This would suggest that perhaps FTDNA obtained a sample that was significantly biased toward faster mutation rates. I am sure that whatever the problem that resulted in FTDNA's rates being higher than Kerchner's or Chandler's, it is likely a subtle one of something like sampling, rather than any mistakes in the handling of the data. It would be very helpful if the FTDNA study could be published.

It is very important that surname project administrators submit their data on their known genealogies to the Kerchner project so that the uncertainties in his rates may be further reduced. This should be done without regard to the number of mutations (or lack of mutations) that have occurred in those projects. There are far more projects having useful data than have been submitted to Kerchner's log so far. For those who have difficulty in understanding how to submit the data, Charles is willing to help. This is another area where our community of "amateurs" is demonstrating that we can make a significant contribution to genetics as applied to genealogy and anthropology.

In a study that uses a known genealogy, there is usually no guesswork necessary in calculating the mutation rate. The number of father-to-son transmissions of the marker set is known, and it is usually possible to reconstruct the haplotype for the common ancestor. Then it becomes a simple matter of counting the mutations observed in the genealogical tree that leads to the present-day participants and dividing by the number of marker transmissions.

However, in many surname projects and in all population studies, the genealogy is not known. This has led to discussions about how to correct for the unknown genealogy, unknown population (or family) dynamics, and the unknown sampling bias that may have been at work in producing the pool of available descendants and the selection of the actual participants.

The number of mutations showing in a group of participants who are all descended from a common ancestor will generally be higher than the actual number of mutations that has occurred in the genealogical history of this group. That is because for some of the mutations presently showing in participants, they will have been inherited by two or more participants from a common ancestor in whom the mutation first appeared. If one simply counts the number of present-day mutations, the derived mutation rate will be too high. If an independent rate is assumed and the TMRCA is calculated, the excess apparent mutations will cause the TMRCA to be too large.

Where the genealogy is not known, there will also be unknown factors of population dynamics at work—some lines from the ancestor will be more prolific than others, biasing the overall results toward the mutation experience of the prolific branch. Other lines may have become extinct. These factors usually have the effect of reducing diversity and causing the calculated TMRCA to be too small. The best way to handle population dynamics is still controversial and the issue is usually ignored.

When FTDNA calculates the TMRCA for a pair of individuals, these issues of genealogy and population dynamics do not apply because the lines from a pair of participants to their most recent common ancestor are (by definition) direct lines with no ambiguities. In this case the father-son mutation rates, rather than the "effective rates," are obviously the appropriate rates to use. However, the results of this calculation will only be as good as the father-son rates that are employed.

Zhivotovsky (2004) published a paper in which he attempted to get around these difficulties by calculating an "effective mutation rate" that is empirically derived from a set of descendants of an ancestor who lived at a known time in the past. All of the unknown factors such as the genealogy or the population dynamics, are simply averaged out in calculating the effective mutation rate, assuming 25 years per generation (which may be too small). This can work well if there are a number of such case studies that can be analyzed and the resulting average rates can be averaged (Zhivotovsky averaged the rates from three population groups), and if the cases that are included are representative of the situation to which the derived rate is to be applied. In practice, it is not so easy to guess whether the case studies have the necessary characteristics to be appropriate.

Zhivotovsky's "effective" mutation rate is averaged over just a few traditionally measured markers. However, Chandler's relative rates for all 37 markers can also be calibrated to Zhivotovsky's effective rates, resulting in a complete set of individual effective rates. Or, any subset of the 37 can be averaged to obtain the corresponding effective average rate for any panel of markers. These would then be appropriate for application to a dataset of haplotypes where the genealogy is unknown and the time depth is similar to that used by Zhivotovsky for his determination of the effective rates.

However, in using three different datasets and averaging the result from each, Zhivotovsky seems to have introduced a small problem: the markers used in the different datasets were not exactly the same, especially for the third dataset, so he was averaging rates over different markers. Even with unlimited sample size, the

rates from the three groups should not be the same. Zhivotovsky averaged them anyway.

However, we can illustrate the approach to recalibrating Chandler's mutation rates to the effective rate by just using the results Zhivotovsky obtained on seven markers that were tested in the dataset of Maori and Cook Islanders, where he obtained a mutation rate averaged over the seven markers of .000705. The seven markers were: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393. If we average Chandler's father-son mutation rates over the same seven markers, we obtain an average value of .00183. Therefore, we can calculate the ratio of Zhivotovsky's effective average rate to Chandler's father-son average rate on those seven markers, as .000705/.00183 = .385. Zhivotovsky's effective rate is just 38.5% of the father-son rate. Armed with that conversion factor, we can convert all 37 of Chandler's individual father-son mutation rates to effective rates *a la* Zhivotovsky by simply multiplying each by 0.385. With this complete set of rates, we can then average them over any subset of markers if desired for a particular application.

It remains rather important that we have an independent check on the mutation rates of Chandler and the average rates of FTDNA. This brings me back to how important it is for individual surname project managers, in cases where the genealogy is known, to submit their data to Charles Kerchner's log (data may be submitted, and results seen, at http://www.ystrlog.org/. I believe that we may soon reach a sufficient amount of data in the log that we could see an article in JoGG on the subject. I have heard a few administrators comment that "it's too complicated [to submit data]," but that just isn't true, though Charles will help if there are questions. As Charles would put it, "Synergy at work!"

Whit Athey

References

Chandler J (2006) Estimating Per-Locus Mutation Rates. *J Genet Geneal*, 2:27-33.

Kerchner (2007) Y-STR Haplotype Observed Mutation Rates in Surname Projects Study and Log, http://www.ystrlog.org/

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G,. Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J Hum Genet, 74:50–61.*