Y-DNA Mutation Rates

A case study of computational models in genetic genealogy

by Dave Hamm
v 1.1
June 28, 2008

The wonderful side effect of using Y-DNA to study Genetic Genealogy is that the DNA begins to crack open long standing problems. One of the most significant problems is calculating the Time to Most Recent Common Ancestor (TMRCA). For the purposes of genetic genealogy, this concept was first devised or proposed in 2001 by Bruce Walsh of the University of Arizona. It has been vexing DNA Project Administrators ever since.

Why is this TMRCA a vexing problem?

Well, the theory proposed by Bruce Walsh dives head first into a perplexing trail of complicated mathematical equations, molecular genetic theory, infinite alleles models, Poisson distributions, Bayesian posterior distributions, Bessel functions, differential mutation rates, and the like. For most Y-DNA administrators, this is a little difficult to apply to their family project(s). Bruce Walsh describes the calculations as an "upper boundary" (page 898) on the time back to a common ancestor shared by two individuals. That is, it was not to be taken as an exact calculation.

The Holy Grail of Genetic Genealogy: Why would people want to know the Time to Most Recent Common Ancestor?

When you begin to use DNA for genealogy, the first thing you are able to determine is which family lines you relate to, and which family lines you do not relate to. That puts a whole new perspective on genealogy research. You begin to understand which family researchers you should be communicating with, and this also leads to which geographical areas should be of interest in your family line. That's because using genetics for TMRCA holds the promise of discovering migration paths going back thousands of years. That is, tracing your own line back several hundred years no longer seems like such a huge problem.

Knowing "How To" calculate the Time to Most Recent Common Ancestor suddenly becomes an interesting problem for genealogists to solve. This calculation should tell you how long it has been since your own line split from another line. Information that genealogists may not have had before. This becomes very interesting as you share resources with other genealogists around the globe.

How TMRCA calculations work

For geneticists, this usually means breaking down probability distributions into a computer program that can be applied to the data. For genealogists, this usually means breaking down those complicated equations into something easy to calculate. For most folks, that means applying a mutation rate. This should be a simple calculation. But, most people find that the "simple" calculation soon breaks down into some type of quagmire when it comes to applying it their own line. The calculations quickly turn into an endless stream of details.

The idea is to apply the mutation rate to the Genetic Distance between your project participants.

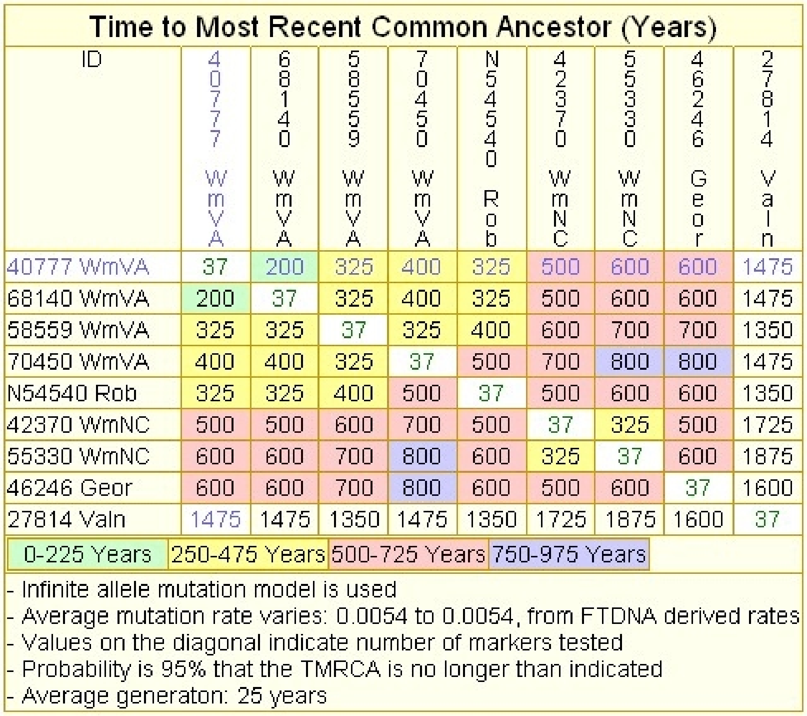
The problem becomes which mutation rate should I apply? Which probability distribution is the correct one? What's the correct mutation model (stepwise, infinite alleles, etc.)? Should I apply individual marker mutation rates to my study? Or, for that matter, how do I figure out the marker mutation rates, and how is that calculated for TMRCA? And even perhaps, how do I know that I have calculated the correct Genetic Distance?

Far too many questions there. Why can't we just point and click on some computer program? I suppose the main reason is that no satisfactory "easy to use" computer program has been written for a genealogist to use in their DNA studies. At this point in time, there is one software program that has attempted something along these lines. That would be Dean McGee's Y-DNA Utility . It at least generates some TMRCA data for you. Dean McGee has made his program available for public use.

One of my favorites, I use the output from Dean McGee's Utility and pass it through the PHYLIP software package to produce a phylogenetic graph of the HAM DNA project . This is all obtained only by the use of the DNA data that we have collected for the project.

Discussion

While I enjoy the output from Dean McGee's Utility, there is a minor problem that I find when applying it to my data. I am observing that individuals with genetic distances of 0, 1, and 2 all descend from a common ancestor who is estimated to have been born in 1755 (or, TMRCA ~= 255 years ago). For these individuals (kits 40777, 68140, 58559, and 70450), Dean McGee's Utility calculates the TMRCA out (for the HAM DNA Project) as 150, 325, and 400 years ago. That is, for each of the differing genetic distances, there are corresponding differing TMRCA's. However, the results should show the same TMRCA. At least, the current McGee output (as given above), which includes 67 marker output. For 37 markers only, Dean McGee's TMRCA output looks more like this:



Time to Most Recent Common Ancestor (Years)

ID	40777 WmVA	68140 WmVA	58559 WmVA	70450 WmVA	N54540 Rob	42370 WmNC	55330 WmNC	46246 Geor	27814 Valn
40777 WmVA	37	200	325	400	325	500	600	600	1475
68140 WmVA	200	37	325	400	325	500	600	600	1475
58559 WmVA	325	325	37	325	400	600	700	700	1350
70450 WmVA	400	400	325	37	500	700	800	800	1475
N54540 Rob	325	325	400	500	37	500	600	600	1350
42370 WmNC	500	500	600	700	500	37	325	500	1725
55330 WmNC	600	600	700	800	600	325	37	600	1875
46246 Geor	600	600	700	800	600	500	600	37	1600
27814 Valn	1475	1475	1350	1475	1350	1725	1875	1600	37

0-225 Years | 250-475 Years | 500-725 Years | 750-975 Years

- Infinite allele mutation model is used
- Average mutation rate varies: 0.0054 to 0.0054, from FTDNA derived rates
- Values on the diagonal indicate number of markers tested
- Probability is 95% that the TMRCA is no longer than indicated
- Average generaton: 25 years

For HAM DNA Group #1, using only 37 marker data gives slightly different output than 67 markers.

Including 67 marker data, TMRCA is reported out as 150, 325, and 400 years ago.
Including only 37 marker data, TMRCA is reported out as 200, 325, and 400 years ago.

The parameters given are

Probability 95 %
FTDNA mutation rate at .004
Infinite alleles model
Generation expressed as 25 years.

If this bothered me to any great extent, I would have contacted Dean McGee directly. But it doesn't. I am quite pleased with the utility. However, this case study is a good example of an extreme error that one might run into using a compute model against such a small number of markers.

The calculations are close to the genealogy information, but not quite exact. How do I resolve that? Can I obtain a TMRCA that actually corresponds better to the actual data? Does this have anything to do with Walsh's "upper boundary" theory? Is there something that Dean McGee's Utility should be doing differently? Are there calculations that I could do on my own data to improve the figure? Or, is it simply due to the number of markers tested? Page 909 of Walsh's paper suggests that per generation data will become accurate when about 580 markers have been tested between two individuals. To my knowledge, only about 417 Y-DNA markers have been discovered, and most testing companies only offer packages of 100 markers or less. Therefore, is the lack of accuracy due to the lack of the number of markers tested?

As for individuals making their own calculations, Bruce Walsh mentions how to modify the TMRCA equations for an individual haplotype (page 910 of his paper).

If I recall correctly, I believe one of Walsh's papers mentioned that Family Tree DNA has at least once considered generating this for their individual projects. But to date, FTDNA has only published information obtained from their data in very general terms, as it applies to the data as a whole.

Therefore, the question follows that if we are able to generate the same type of information from our own project (or haplotype), then will that data result in a more accurate estimation of TMRCA as it applies to our area of interest?

As of this writing, there are at least two individuals that have published thoughts along those lines (at least, to my knowledge). One is Charles Kerchner, who is tracking data across a multitude of projects. Charles is studying mutation rates for individual projects.

The other individual is David Roper, who shows how he has applied calculations for his own project on a very small scale. Mr. Roper has included some discussion of how to apply probabilities to genetic distance for an individual project , and he has posted the results of a simple example of "How To" calculate this out.

Comparison of Calculations:

The calculations were compared to a special case in the HAM DNA Project, where no mutations were observed within a TMRCA of 250 years.

The Roper Model:

David Roper had reduced Walsh's equations (see equation 12) down to a calculation such as:

(i.e., Roper ref: Walsh)

2 * mu * t = ln[n/k]

where n markers with k matches and mu = 1/500

t = 250 * ln [n/k] in generations.

But, this natural log equation is not the example that I want to focus upon here. (I will briefly comment on this equation at the end of this paper.)

Roper's approach is interesting, because he calculates out the "variability" of individual markers in a standard statistical manner. He then applies the variability to obtain a "probability" figure. Then, using genetic distance he is able to derive a TMRCA. Once you step through this, you will begin to appreciate a calculator or a software program. It should be noted that this is not exactly the computation that Walsh suggests (for small projects), but it is a simplified model that can be applied by genetic genealogists.

Roper calculates (from the data in his project for a case study):

- variability as the average difference in values per marker (avg 4.56 - found empirically on his own from Project data)
- Probability as (number of markers * mutation rate / variability)

= 25/500/4.56
= .01096 per marker

which gives:

Variability per marker	1	2	3	4	5	6	7	8	9
Probability per marker	0.0110	0.0219	0.0329	0.0439	0.0548	0.0658	0.0768	0.0877	0.0987

He applies these probability values per marker in order to obtain a per marker genetic distance. Then, he adds up the per marker genetic distance in order to obtain an estimate of the number of generations per marker.

For example, he gets a genetic distance of "1" from DYS385a.
He then divides that by .0877 to get 11.4 in generations (represented by DYS385a).

1 / .0877 = 11.4

Summing all of his markers values for generations together, he gets:

11.4 + 11.4 + 36.5 + 18.2 + 30.4 + 20.3 + 15.2 + 18.2 = 161.6 total generations (times 2), or 80.80 generations

By using the variability of his own data, he obtains a TMRCA of 2,000 years in contrast to using a standard mutation rate that would otherwise calculate out to 2,500 years (in his simple example). That roughly corresponds to a difference in mutation rate of 1/500 (.0002) vs. a mutation rate of 1/250 (.0004).

However, you might notice that Roper roughly corresponds to what Bruce Walsh had published as a "standard" mutation rate as compared to a "high" mutation rate in his " Technical Details for TMRCA Calculations ." (The mutation rate of .0002 is the average of a number of studies, .0004 is the likely "under" estimate of the true time.)

Dean McGee's Utility

I mentioned the slight problem with Dean McGee's Utility for the individuals (kits 40777, 68140, 58559, and 70450). These kits are contrasted by two markers, DYS439 and CDYb.
By the following genetic distance per marker:

	DYS439	CDYb
40777	0	0
68140	0	0
58559	0	1
70450	1	1

The genetic distance is the same for this group, as there have been no mutations observed between 37 and 67 markers for this group.

For 67 markers, Dean McGee's Utility reports this case study out as:

Kit numbers	**Genetic Distance**	**Actual TMRCA**	**McGee TMRCA**
	Observed	**(from genealogy research)**	**(95% at .004 inf. alleles from Y-DNA Utility)**
40777 & 68140	**0**	**10 generations (or est. 255 years)**	**0 generations (or est. 150 yrs)**
40777 & 58559	**1 (DYS439)**	**10 generations (or est. 255 years)**	**13 generations (or est. 325 yrs)**
40777 & 70450	**2 (DYS439 & CDYb)**	**10 generations (or est. 255 years)**	**16 generations (or est. 400 yrs**)

Or roughly, and error of

[(255 - 150) + (325 - 255) + (400 - 255))]/3 = (105 + 70 + 145)/3 = 106.7 years

For 37 markers, Dean McGee's Utility reports this case study out as:

Kit numbers	**Genetic Distance**	**Actual TMRCA**	**McGee TMRCA**
	Observed	**(from genealogy research)**	**(95% at .004 inf. alleles from Y-DNA Utility)**
40777 & 68140	**0**	**10 generations (or est. 255 years)**	**8 generations (or est. 200 yrs)**
40777 & 58559	**1 (DYS439)**	**10 generations (or est. 255 years)**	**13 generations (or est. 325 yrs)**
40777 & 70450	**2 (DYS439 & CDYb)**	**10 generations (or est. 255 years)**	**16 generations (or est. 400 yrs**)

Or roughly, 37 markers reports an error of

[(255 - 200) + (325 - 255) + (400 - 255))]/3 = (55 + 70 + 145)/3 = 90.0 years

It is counter intuitive that 37 markers would report out better TMRCA estimates than 67 markers would for this group. At first glance, I would suppose this would be due to not yet having exact mutation rates for the markers in the 38 to 67 marker range (as of June, 2008). However, that reasoning would not be correct, since there are no mutations to report for this group between markers 38 to 67.

Therefore, it is more reasonable to presume that a "baseline" value for the time estimate is required when no mutations are reported. This is interesting because such a "baseline" would appear to be dependent upon the number of markers reported, and not simply a constant value.

Lamarc output:

(Lamarc - Likelihood Analysis with Metropolis Algorithm using Random Coalescense)

It should be noted that the Lamarc output could be applied very differently than I give here. The Lamarc documentation provides some information about applying Theta to haploids, diploids, and the special case where gene is only passed from father to son (such as Y-DNA).

My example here simply follows that if Theta is proportional to the mutational rate, then the proportion of the individual marker Theta values to the overall Group Theta values should be proportional to the mutation rates for these values.

Group #1 Theta MLE for DYS439 = 0.018819
Group #1 Theta MLE for CDYb = 0.024394

Group #1 Theta MLE overall = 0.020130

DYS439 has a Theta value of slightly lower than the Theta for the group overall (93.48 %).
CDYb has a Theta value of slightly higher than the Theta for the group overall (121.2 %)

1 / .018819 = 53.14
1 / .024394 = 40.99

53.14 x .02013 = 1.07 (genetic distance due to DYS439)
40.99 x .02013 = 0.825 (genetic distance due to CDYb)

If Theta is proportional to the mutation rate, then comparing the values for Theta should show the relative proportion to the mutation rate.

If the proportion of Theta to the mutation rate is a constant, then:

0.018819 / 0.020130 = .9348 (ratio of the mutation rate for DYS439 in Group #1)
0.024394 / 0.020130 = 1.2112 (ratio of the mutation rate for CDYb in Group #1)

That is, if the mutation rate is .004, then:

.004 x .9348 = .00374 (mutation rate for DYS439 in Group #1)
.004 x 1.2112 = .00485 (mutation rate for CDYb in Group #1)

given a mutation rate of .004

Applying a mutation rate of .004 gives

mutation rate of DYS439
1 / .00374 = 267.4 years
mutation rate of CDYb
1 / .00485 = 206.2 years

straight addition would give a genetic distance of DYS439 + CDYb = 267.4 + 206.2 = 473.6 years (times 2)

Or, (267.4 + 206.2) / 2 = 236.8 years (i.e., born ~ est 2005 - 236.8 = 1768)

In this case, the average for three individuals (including the genetic distance of "0") only makes sense if there are 35 years per generation.

Kit numbers	**Genetic Distance Observed**	**Actual TMRCA (from genealogy research)**	**Lamarc TMRCA (.004 mutation rate from marker theta)**
40777 & 68140	**0**	**10 generations (or est. 255 years)**	**0 generations (or est. 0 years)**
40777 & 58559	**1 (DYS439)**	**10 generations (or est. 255 years)**	**10.7 generations (or est. 267.4 years)**
40777 & 70450	**2 (DYS439 & CDYb)**	**10 generations (or est. 255 years)**	**9.5 generations (or est. 236.8 years)**

McGee's Utility (67 markers, infinite alleles model at 95%) is off by (105 + 70 + 145) / 3 = 107 years on avg

McGee's Utility (37 markers, infinite alleles model at 95%) is off by (55 + 70 + 145) / 3 = 90 years on avg

Lamarc Theta ratios with mutation rate of .004 is off by (250 + 17.4 + 13) /3 = 93.5 years on average, suggesting a baseline of 156.5 years.

That is, if the mutation rate is .002, then the Lamarc example becomes:

0.018819 / 0.020130 = .9348 (ratio of the mutation rate for DYS439 in Group #1)
0.024394 / 0.020130 = 1.2112 (ratio of the mutation rate for CDYb in Group #1)

.002 x .9348 = .00187 (mutation rate for DYS439 in Group #1)
.002 x 1.2112 = .00242 (mutation rate for CDYb in Group #1)

given a mutation rate of .002

Applying a mutation rate of .002 gives

genetic distance of DYS439
1 / .00187 = 534.8
genetic distance of CDYb
1 / .00242 = 413.2

straight addition would give a genetic distance of DYS439 + CDYb = 534.8 + 413.2 = 948 (times 2)

Or, (534.8 + 413.2) / 2 = 474 years (i.e., born ~ est 2005 - 474 = 1531)

In this case, the average for three individuals (including the genetic distance of "0") does not make sense.

Lamarc Theta ratios with mutation rate of .002 is off by (250 + 285 + 224) /3 = 253 years on average

----

Therefore, the appropriate method is to use the Lamarc theta ratios with the mutation rate of .004, since the mutation rate of .002 does not appear to be realistic at all. Also, the Lamarc ratios make no valid estimates when there are no mutations observed. That is, it may be best to use the Lamarc ratios with some baseline in order to account for the special case of no mutating markers. For example, if no mutations are observed, the Lamarc baseline could be set to (250 - 93.5 = 156.5 years ago, or 6 generations ago).

For example, if a baseline of 156.5 years is applied to the Lamarc data when no mutations are observed, the values become:

Kit numbers	**Genetic Distance**	**Actual TMRCA**	**Lamarc TMRCA**
	Observed	**(from genealogy research)**	**(.004 mutation rate from marker theta)**
40777 & 68140	**0**	**10 generations (or est. 255 years)**	**6.3 generations (or est. 156.5 yrs)**
40777 & 58559	**1 (DYS439)**	**10 generations (or est. 255 years)**	**10.7 generation (or est. 267.4 yrs)**
40777 & 70450	**2 (DYS439 & CDYb)**	**10 generations (or est. 255 years)**	**9.5 generations (or est. 236.8 yrs)**

McGee's Utility (67 markers, infinite alleles model at 95%) is off by (105 + 70 + 145) / 3 = 107 years on avg
McGee's Utility (37 markers, infinite alleles model at 95%) is off by (55 + 70 + 145) / 3 = 90 years on avg
Lamarc Theta ratios with mutation rate of .004 is off by (250 + 17.4 + 13) /3 = 93.5 years on avg, suggesting a baseline of 156.5 years.
Lamarc Theta ratios with mutation rate of .004 and baseline of 156.5 years is off by (93.5 + 17.4 + 13) /3 = 41.3 yrs avg.

Using a baseline with the Lamarc data *doubles the accuracy of the model*.

As of 06/27/2008, for the HAM DNA Project:

Roper probability model using a mutation rate of 0.002:

- variability as the average difference in values per marker (avg 1.33) - as of 06/26/2008 - 12 / 9 = 1.33
- Probability as (number of markers * mutation rate / variability)
= 37/500/1.33
= .0556 per marker

which gives:

Variability	1	2	3	4	5	6	7	8	9
Probability	0.0556	0.1113	0.1669	0.22256	0.2782	0.33383	0.3895	0.4451	0.50075

DYS439 has a variable value of 1 = 0.0556
CDYb has a variable value of 2 = 0.1113

for a genetic distance of 1 on DYS439, 1 / 0.0556 = 17.986 generations
for a genetic distance of 1 on CDYb, 1 / 0.113 = 8.85 generations

17.986 + 8.85 = 26.836 total generations (times two)

or, DYS439 + CDYb gives total generations of 26.836 / 2 = 13.418 (or 25 x 13.418 = 335 years)

DYS439 alone gives total generations of 8.85 / 2 = 4.425 (or 25 x 4.425 = 110.6 years)

Kit numbers	**Genetic Distance**	**Actual TMRCA**	**Roper TMRCA**
	Observed	**(from genealogy research)**	**(1.33 at .004 from marker variability)**
40777 & 68140	**0**	**10 generations (or est. 255 years)**	**0 generations (or est. 0 yrs)**
40777 & 58559	**1 (DYS439)**	**10 generations (or est. 255 years)**	**1 generation (or est. 111 yrs)**
40777 & 70450	**2 (DYS439 & CDYb)**	**10 generations (or est. 255 years)**	**2 generations (or est. 335 yrs)**

McGee's Utility (67 markers, infinite alleles model at 95%) is off by (105 + 70 + 145) / 3 = 107 years on avg
McGee's Utility (37 markers, infinite alleles model at 95%) is off by (55 + 70 + 145) / 3 = 90 years on avg
Lamarc Theta ratios with mutation rate of .004 is off by (250 + 17.4 + 13) /3 = 93.5 years on avg, suggesting a baseline of 156.5 years.
Lamarc Theta ratios with mutation rate of .004 and baseline of 156.5 years is off by (93.5 + 17.4 + 13) /3 = 41.3 yrs avg.
Roper's model (mutation rate of .002) is off by (250 + 139 + 85) / 3 = 158 years on average

----

As of 06/24/2008, for the HAM DNA Project Group #1:

Roper probability model using a mutation rate of 0.004:

- variability as the average difference in values per marker (avg 1.33)
as of 06/27/2008 - 12 / 9 = 1.33
- Probability as (number of markers * mutation rate / variability)

= 37/250/1.33
= .1113 per marker

which gives:

Variability	1	2	3	4	5	6	7	8	9
Probability	0.1113	0.22256	0.33383	0.4451	0.5564	0.6677	0.779	0.89023	1.0015

DYS439 has 1 variable value of 1 = 0.1113
CDYb has 2 variable values of 1 = 0.22256

for a genetic distance of 1 on DYS439, 1 / 0.1113 = 8.99 generations
for a genetic distance of 1 on CDYb, 1 / 0.22256 = 4.49 generations

8.99 + 4.49 = 13.48 total generations (times two)

or, DYS439 + CDYb gives total generations of 13.48 / 2 = 6.74 (or 25 x 6.74 = 168.5 years)

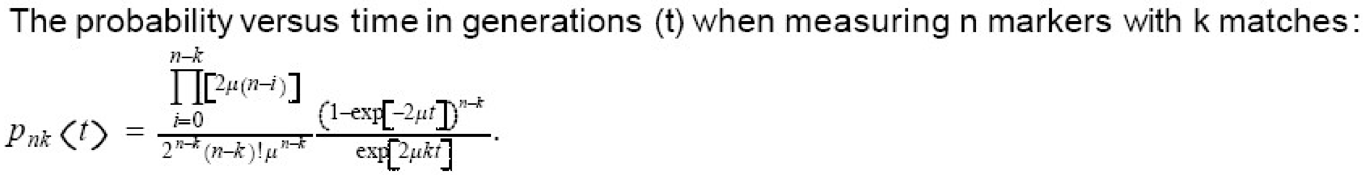
DYS439 alone gives total generations of 8.99 / 2 = 4.49 (or 25 x 4.49 = 112.4 years)

Kit numbers	**Genetic Distance**	**Actual TMRCA**	**Roper model TMRCA**
	Observed	**(from genealogy research)**	**(1.33 at .004 from marker variability)**
40777 & 68140	**0**	**10 generations (or est. 255 years)**	**0 generations (or est. 0 years)**
40777 & 58559	**1 (DYS439)**	**10 generations (or est. 255 years)**	**1 generation (or est. 112 years)**
40777 & 70450	**2 (DYS439 & CDYb)**	**10 generations (or est. 255 years)**	**2 generations (or est. 337 years)**

McGee's Utility (67 markers, infinite alleles model at 95%) is off by (105 + 70 + 145) / 3 = 107 years on avg
McGee's Utility (37 markers, infinite alleles model at 95%) is off by (55 + 70 + 145) / 3 = 90 years on avg
Lamarc Theta ratios with mutation rate of .004 is off by (250 + 17.4 + 13) /3 = 93.5 years on avg,
suggesting a baseline of 156.5 years.
Lamarc Theta ratios with mutation rate of .004 and baseline of 156.5 years is off by (93.5 + 17.4 + 13) /3 = 41.3 yrs avg.
Roper's model (mutation rate of .002) is off by (250 + 139 + 85) / 3 = 158 years on average
Roper's model (mutation rate of .004) is off by (250 + 138 + 87) / 3 = 158 years on average
Lamarc Theta ratios with mutation rate of .004 and baseline of 156.5 years is off by (93.5 + 17.4 + 13) /3 = 41.3 years on average.

Getting back to David Roper's reduced form of Walsh's equation:

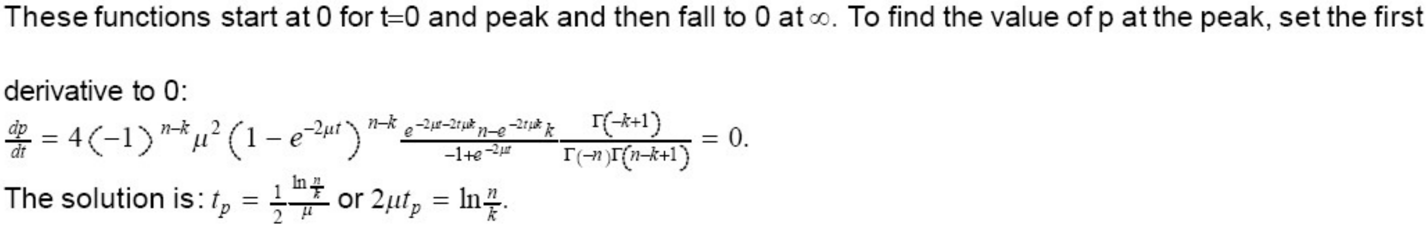
Basically, Roper reduced Walsh's equation (see equation 12) from this:



The probability versus time in generations (t) when measuring n markers with k matches:

$$p_{nk}\langle t\rangle = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\frac{(1-\exp[-2\mu t])^{n-k}}{\exp[2\mu k t]}.$$

which Roper reduced to this:



These functions start at 0 for t=0 and peak and then fall to 0 at ∞. To find the value of p at the peak, set the first derivative to 0:

$$\frac{dp}{dt} = 4(-1)^{n-k}\mu^2\left(1-e^{-2\mu t}\right)^{n-k}\frac{e^{-2\mu t-2t\mu k}n-e^{-2t\mu k}k}{-1+e^{-2\mu t}}\frac{\Gamma(-k+1)}{\Gamma(-n)\Gamma(n-k+1)} = 0.$$

The solution is: $t_p = \frac{1}{2}\frac{\ln\frac{n}{k}}{\mu}$ or $2\mu t_p = \ln\frac{n}{k}$.

(i.e., Roper ref: Walsh)

2 * mu * t = ln[n/k]

where n markers with k matches and mu = 1/500

t = 250 * ln [n/k] in generations

This equation appears to be best applied as Roper applied it. That is, in terms of comparing matching markers per individual, out of the total number of markers tested:

where n markers with k matches and mu = 1/500

$t = 250 * \ln [n/k]$ in generations

For 37 markers, n = 37

$k = 37$: $t = 250 \ln [37/37] = 0.0$
$k = 36$: $t = 250 \ln [37/36] = 6.8497$
$k = 35$: $t = 250 \ln [37/35] = 13.892$
$k = 34$: $t = 250 \ln [37/34] = 21.139$

Which is not entirely accurate for the specific case that I examine here (kits 40777, 68140, 58559, and 70450).

In this case study, I am looking at a very small population where most markers do match within a given group. These restrictions would affect the way in which Walsh's equations should be derived. Which is to say, I should be using a slightly different equation for what I am examining here.

Put simply, both Walsh's equations and the data suggest that I should be using a baseline value and/or a different equation for small population studies (such as Group #1 of the HAM DNA Project).However, it is these small population studies that interest genetic genealogists the most.

Observations

Many of the genetic genealogists are looking for specific results as they apply to their own project or group. They are not always content with generalized information, and the question of TMRCA is a topic of great interest especially if they can apply an equation to their own project(s).

It should be noted that a significant percentage of mutations have been reported for father son pairs. I have not examined the effect of a baseline to the case of father son pairs (mainly because I do not have that data at hand).

I should also note that there are any number of genetic software programs available, usually requiring input in the form of ATGC format. Nearly all of the genetic programs do not report output that the family genealogist can easily use to compute a reasonable TMRCA. That is, most programs do not take the input data as we have it from FTDNA, nor do most programs report TMRCA output expressed in terms such as generations or years. (See Bill Jackson's MRCA Probability calculator for a given number of markers and generations, or Ann Turner's Mutation calculator.)

It should also be noted that other quantized methods could be employed to calculate meaningful results (even without the use of a mutation rate). However to date, I have not yet examined other quantized computational models that could be employed by geneticists for TMRCA.

I have not derived the appropriate equations (for a small population) from Bruce Walsh's paper. (The mathematics appear to be beyond my abilities.) Also, there would be more appropriate use of the output from Lamarc, I used a simple example of proportions here for convenience.

Finally, it should be remembered that this study is based upon a special (and rather unusual) case of 37 marker computations. That means, due to the low number of markers examined, probabilities are a large factor in the results. Last I checked, there have been about 4317 Y-DNA markers discovered, bringing to mind Hammer's comment to Walsh (in Walsh's paper) that about 580 Y-DNA markers would be required to resolve time frames down to each generation. When we finally have good analysis on a large number of markers, then this type of discussion will probably become a non issue.

Conclusions

Since most participants have not gone through the process of detailed calculations for per marker mutation rates (for individual projects), it has yet to be determined if all of the work is useful. It is interesting that Roper has detailed the probability calculations for individual markers, which differed slightly from the "standard" calculations using a "standard" mutation rate. This type of research could be useful for individual projects.

How were the models checked independently for the HAM DNA Project?

An attempt was made to calculate TMRCA from various independent computational models. First off, I am presuming that my methodology and math is reasonably correct. There are a number of opportunities to introduce errors in any of the above computations. For example, using the Lamarc program could introduce errors when the data from FTDNA is translated into ATGC format. Or, more simply, my math in this report may not have been applied appropriately.

I had enough data for a Lamarc run on my project for two groups, Group #1 and Group #2. Lamarc calculates out "Theta" values for the Group as a whole, and also Theta values per individual marker. This appears to generate slightly more accurate TMRCA estimates than Dean McGee's Y-DNA Utility. (Roughly, a 13 % improvement in TMRCA estimates.) However, the Lamarc program can run for several days, and takes a great deal more effort than does Dean McGee's Utility.

Finally, it was found that adding a baseline value for no mutations doubles the accuracy of the Lamarc data. Of course, a baseline could also be applied to Dean McGee's utility, or to Roper's model as well. Therefore, one can only wonder how accurate Bruce Walsh's equations might be if a simple baseline were added to his equations.

At the moment, Dean McGee's Y-DNA Utility appears to provide the best calculations with considerably less effort. Also, the Y-DNA Utility can be adjusted for model (infinite alleles or hybrid), for probability, and mutation rate. I would have to conclude that Dean McGee's Utility is the best program available for ease of use, parameters offered, and of course, price.

See Also:

Time to Most Recent Common Ancestor (TMRCA) a PDF file by Bruce Walsh (2001) of the University of Arizona.

Dienke's Anthropology blog regarding Y-DNA mutation rates of father-son pairs (posted in 2006).

Dean McGee's Y-DNA Utility

David Roper's discussion of how to apply probabilities to genetic distance for an individual project , and his posted results

PHYLIP software package

Mutation Rate calculations by Rosche and Foster (2006)

Lamarc - Likelihood Analysis with Metropolis Algorithm using Random Coalescence