HAM DNA Project

TMRCA calculation model
using Lamarc MLE output for Group #02

by Dave Hamm

July 4, 2008

In an effort to see if there may be a better compute model for the Y-DNA data for the HAM DNA Project, an example was run through the LAMARC Program (Maximum Likelihood Parameter Estimation using Hastings-Metropolis Markov Chain Monte Carlo).

The Lamarc program started on 10/13/07 10:41:40 and finished on 10/15/07 15:22:39
Nine kits were used from the HAM DNA Project, 36 markers (or 36 regions) were used, totaling 324 samples in all regions. The conversion from FTDNA data to ATGC format was done with the "Ft2Dna" program.

The Lamarc output produced MLE Theta values for Group #02 overall, as well for the existing mutating markers within Group #02. There were 8 mutating markers for this group.

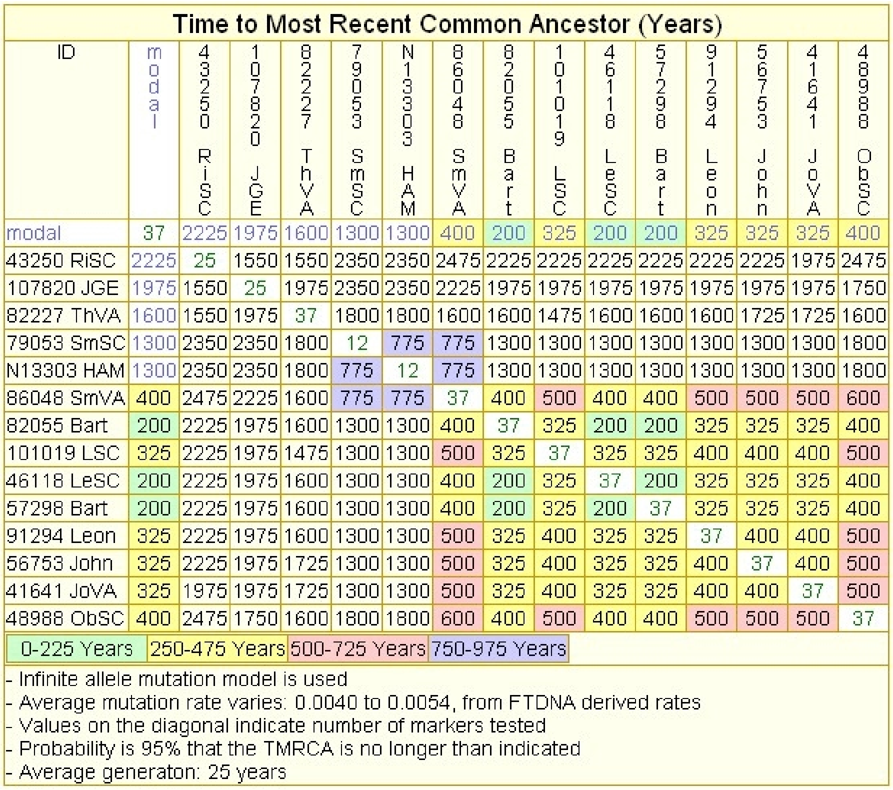
The results were then converted from Theta values into a corresponding mutation rate for the group and each mutating marker. This conversion was based upon the mutation rate suggested by Family Tree DNA (FTDNA) at .004.

Dean McGee's Y-DNA Comparison Utility was then run on the existing data for 37 markers in the Group. The output from the conversion of the Theta values from Lamarc into individual marker mutation rates was then compared to the output from Dean McGee's Y-DNA Utility for TMRCA.

The results were comparable to Dean McGee's Utility.

One kit (#56753) was notably re-arranged on a corresponding TMRCA phylogenetic graph, in comparison to the normal phylogram produced from the output given by Dean McGee's Utility.

The normal output from Dean McGee's Y-DNA Utility is given below:



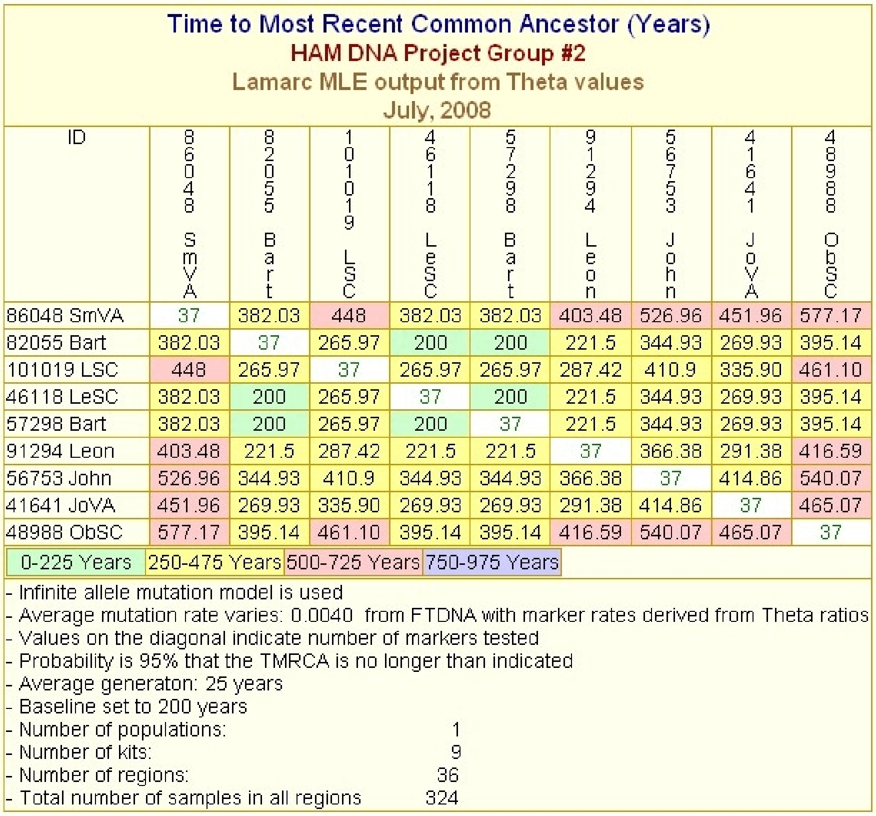
Time to Most Recent Common Ancestor (Years)

ID	modal	43250 RiSC	107820 JGE	82227 ThVA	79053 SmSC	N13303 HAM	86048 SmVA	82055 Bart	101019 LSC	46118 LeSC	57298 Bart	91294 Leon	56753 John	41641 JoVA	48988 ObSC
modal	37	2225	1975	1600	1300	1300	400	200	325	200	200	325	325	325	400
43250 RiSC	2225	25	1550	1550	2350	2350	2475	2225	2225	2225	2225	2225	2225	1975	2475
107820 JGE	1975	1550	25	1975	2350	2350	2225	1975	1975	1975	1975	1975	1975	1975	1750
82227 ThVA	1600	1550	1975	37	1800	1800	1600	1600	1475	1600	1600	1600	1725	1725	1600
79053 SmSC	1300	2350	2350	1800	12	775	775	1300	1300	1300	1300	1300	1300	1300	1800
N13303 HAM	1300	2350	2350	1800	775	12	775	1300	1300	1300	1300	1300	1300	1300	1800
86048 SmVA	400	2475	2225	1600	775	775	37	400	500	400	400	500	500	500	600
82055 Bart	200	2225	1975	1600	1300	1300	400	37	325	200	200	325	325	325	400
101019 LSC	325	2225	1975	1475	1300	1300	500	325	37	325	325	400	400	400	500
46118 LeSC	200	2225	1975	1600	1300	1300	400	200	325	37	200	325	325	325	400
57298 Bart	200	2225	1975	1600	1300	1300	400	200	325	200	37	325	325	325	400
91294 Leon	325	2225	1975	1600	1300	1300	500	325	400	325	325	37	400	400	500
56753 John	325	2225	1975	1725	1300	1300	500	325	400	325	325	400	37	400	500
41641 JoVA	325	1975	1975	1725	1300	1300	500	325	400	325	325	400	400	37	500
48988 ObSC	400	2475	1750	1600	1800	1800	600	400	500	400	400	500	500	500	37

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

- Infinite allele mutation model is used
- Average mutation rate varies: 0.0040 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

In comparison, when the Lamarc conversion was made, it produced the following table:



Time to Most Recent Common Ancestor (Years)

HAM DNA Project Group #2

Lamarc MLE output from Theta values

July, 2008

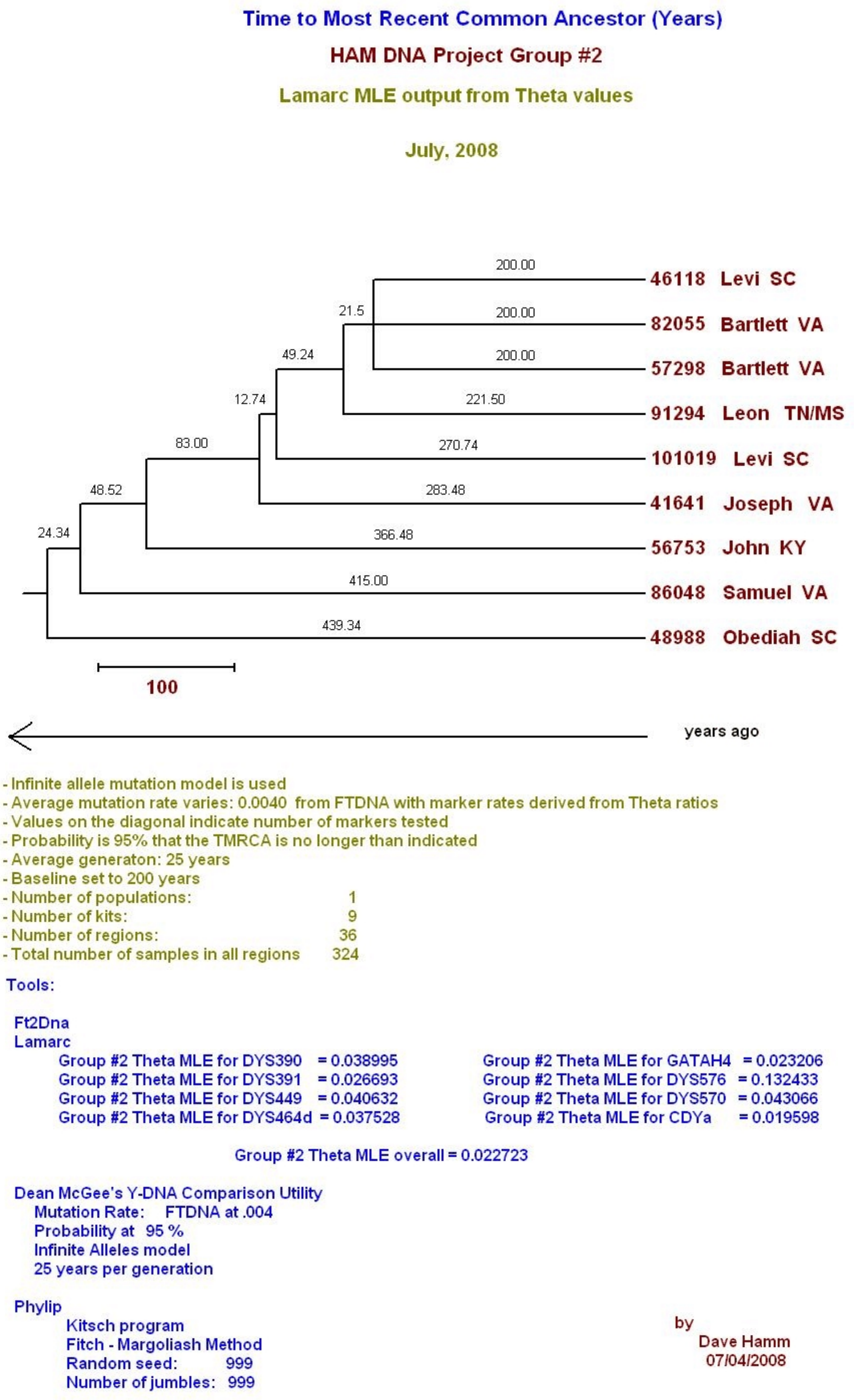
ID	86048 SmVA	82055 Bart	101019 LSC	46118 LeSC	57298 Bart	91294 Leon	56753 John	41641 JoVA	48988 ObSC
86048 SmVA	37	382.03	448	382.03	382.03	403.48	526.96	451.96	577.17
82055 Bart	382.03	37	265.97	200	200	221.5	344.93	269.93	395.14
101019 LSC	448	265.97	37	265.97	265.97	287.42	410.9	335.90	461.10
46118 LeSC	382.03	200	265.97	37	200	221.5	344.93	269.93	395.14
57298 Bart	382.03	200	265.97	200	37	221.5	344.93	269.93	395.14
91294 Leon	403.48	221.5	287.42	221.5	221.5	37	366.38	291.38	416.59
56753 John	526.96	344.93	410.9	344.93	344.93	366.38	37	414.86	540.07
41641 JoVA	451.96	269.93	335.90	269.93	269.93	291.38	414.86	37	465.07
48988 ObSC	577.17	395.14	461.10	395.14	395.14	416.59	540.07	465.07	37

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

- Infinite allele mutation model is used
- Average mutation rate varies: 0.0040 from FTDNA with marker rates derived from Theta ratios
- Values on the diagonal indicate number of markers tested
- Probability is 95% that the TMRCA is no longer than indicated
- Average generaton: 25 years
- Baseline set to 200 years
- Number of populations: 1
- Number of kits: 9
- Number of regions: 36
- Total number of samples in all regions 324

The data was then run through the Phylip package, running the Kitsch program, and using the Fitch-Margoliash method. A TMRCA phylogram was then produced.

The resulting TMRCA phylogram by use of the Lamarc data looked like this:



Time to Most Recent Common Ancestor (Years)

HAM DNA Project Group #2

Lamarc MLE output from Theta values

July, 2008

- Infinite allele mutation model is used
- Average mutation rate varies: 0.0040 from FTDNA with marker rates derived from Theta ratios
- Values on the diagonal indicate number of markers tested
- Probability is 95% that the TMRCA is no longer than indicated
- Average generaton: 25 years
- Baseline set to 200 years
- Number of populations: 1
- Number of kits: 9
- Number of regions: 36
- Total number of samples in all regions 324

Tools:

Ft2Dna
Lamarc

Group #2 Theta MLE for DYS390 = 0.038995	Group #2 Theta MLE for GATAH4 = 0.023206
Group #2 Theta MLE for DYS391 = 0.026693	Group #2 Theta MLE for DYS576 = 0.132433
Group #2 Theta MLE for DYS449 = 0.040632	Group #2 Theta MLE for DYS570 = 0.043066
Group #2 Theta MLE for DYS464d = 0.037528	Group #2 Theta MLE for CDYa = 0.019598

Group #2 Theta MLE overall = 0.022723

Dean McGee's Y-DNA Comparison Utility
Mutation Rate: FTDNA at .004
Probability at 95 %
Infinite Alleles model
25 years per generation

Phylip
Kitsch program
Fitch - Margoliash Method
Random seed: 999
Number of jumbles: 999

by
Dave Hamm
07/04/2008

Details on the calculations follow below.

Group #2 had a number of closely matching participants that have tested for 37 markers.
The following is a table of mutating the markers for Group #2:

kit	DYS390	DYS391	DYS449	DYS464d	GATAH4	DYS576	DYS570	CDYa
reference modal	24	12	31	17	11	17	16	36
86048	0	1	0	1	0	0	0	0
46118	0	0	0	0	0	0	0	0
101019	0	0	0	0	0	0	1	0
82055	0	0	0	0	0	0	0	0
57298	0	0	0	0	0	0	0	0
91294	0	0	0	0	0	1	0	0
56753	0	0	0	0	0	0	0	1
41641	0	0	1	0	0	0	0	0
48988	1	0	0	0	1	0	0	0

Note that these are very small differences, but they vary widely amongst the different markers.
As a group, the genetic distance has a maximum of 4 (between 86048 and 48988).

```
------------------------
From Lamarc:

Number of populations:                          1
Number of regions:                             36
Total number of samples in all regions        324
-----------------------
```

Group #2 Theta MLE for DYS390 = 0.038995
Group #2 Theta MLE for DYS391 = 0.026693
Group #2 Theta MLE for DYS449 = 0.040632
Group #2 Theta MLE for DYS464d = 0.037528
Group #2 Theta MLE for GATAH4 = 0.023206
Group #2 Theta MLE for DYS576 = 0.132433
Group #2 Theta MLE for DYS570 = 0.043066
Group #2 Theta MLE for CDYa = 0.019598

Group #2 Theta MLE overall = 0.022723

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:

0.038995 / 0.022723 = 1.7161 (ratio of the mutation rate for DYS390 in Group #2)
0.026693 / 0.022723 = 1.1747 (ratio of the mutation rate for DYS391 in Group #2)
0.040632 / 0.022723 = 1.7881 (ratio of the mutation rate for DYS449 in Group #2)
0.037528 / 0.022723 = 1.6515 (ratio of the mutation rate for DYS464d in Group #2)
0.023206 / 0.022723 = 1.0213(ratio of the mutation rate for GATAH4 in Group #2)
0.132433 / 0.022723 = 5.8282 (ratio of the mutation rate for DYS576 in Group #2)
0.043066 / 0.022723 = 1.8953 (ratio of the mutation rate for DYS570 in Group #2)
0.019598 / 0.022723 = 0.8625 (ratio of the mutation rate for CDYa in Group #2)

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

.004 x 1.7161 = .00686 (mutation rate for DYS390 in Group #2)
.004 x 1.1747 = .00470 (mutation rate for DYS391 in Group #2)
.004 x 1.7881 = .00715 (mutation rate for DYS449 in Group #2)
.004 x 1.6515 = .00661 (mutation rate for DYS464d in Group #2)
.004 x 1.0213 = .00409 (mutation rate for GATAH4 in Group #2)
.004 x 5.8282 = .02331 (mutation rate for DYS576 in Group #2)
.004 x 1.8953 = .00758 (mutation rate for DYS570 in Group #2)
.004 x 0.8625 = .00345 (mutation rate for CDYa in Group #2)

Applying a mutation rate of .004 gives

mutation rate of DYS390
1 /.00686 = 145.77 years

mutation rate of DYS391
1 /.00470 = 212.77 years

mutation rate of DYS449
1 /.00715 = 139.86 years

mutation rate of DYS464d
1 /.00661 = 151.29 years

mutation rate of GATAH4
1 /.00409 = 244.50 years

mutation rate of DYS576
1 /.02331 = 42.90 years

mutation rate of DYS570
1 /.00758 = 131.93 years

mutation rate of CDYa
1 /.00345 = 289.86 years

I was not able to determine the equation that I should use as a baseline for the data.

Given that 82055 and 57298 and 46118 all match on 37 markers, but the MRCA is not yet known.

McGee's Utility puts an average baseline TMRCA of 200 years for these three (for 37 markers).
Bruce Walsh's 897 paper gives (for small populations):

lambda = 1 / Ne

Lamarc output has 324 samples, 9 kits were used, 8 markers are mutating for this group.

The following are some sample calculations:

----
between 56753 and any of the three (82055 or 57298 or 46118)
off by CDYa = 289.86 years (times two)

Or, 289.86 / 2 = 144.93 years
adding in a baseline of 200 gives: 344.93 years
comparing to McGee's Utility estimate of 325 years
----
between 86048 and any of the three (82055 or 57298 or 46118)
off by DYS391 and DYS464d
DYS391 = 212.77 years (times two)
DYS464d = 151.29 years (times two)

Or, (212.77 + 151.29) / 2 = 182.03 years
adding in a baseline of 200 gives: 382.03 years
comparing to McGee's Utility estimate of 400 years
----
between 48988 and any of the three (82055 or 57298 or 46118)
off by DYS390 and GATAH4
DYS390 = 145.77 years (times two)
GATAH4 = 244.50 years (times two)

Or, (145.77 + 244.50) / 2 = 195.14 years
adding in a baseline of 200 gives: 395.14 years
comparing to McGee's Utility estimate of 400 years
----
between 56753 and 41641
off by CDYa and DYS449
(289.86 + 139.86) = 429.72 years (times two)

Or, 429.72 / 2 = 214.86 years
adding in a baseline of 200 gives: 414.86 years
comparing to McGee's Utility estimate of 400 years
----

between 56753 and 101019
off by CDYa and DYS570
= 289.86 + 131.93
= 421.79 years (times two)

Or, 421.79 / 2 = 210.9 years
adding in a baseline of 200 gives: 410.9 years
comparing to McGee's Utility estimate of 400 years

----

between 56753 and 91294
off by CDYa and DYS576
= 289.86 + 42.90
= 332.76 years (times two)

Or, 332.76 / 2 = 166.38 years
adding in a baseline of 200 gives: 366.38 years
comparing to McGee's Utility estimate of 400 years

----

between 56753 and 48988
off by CDYa and DYS390 and GATAH4
= 289.86 + 145.77 + 244.50
= 680.13 years (times two)

Or, 680.13 / 2 = 340.07 years
adding in a baseline of 200 gives: 540.07 years
comparing to McGee's Utility estimate of 500 years

----

between 91294 and 101019
off by DYS576 and DYS570
= 42.90 + 131.93
= 174.83 years (times two)

Or, 174.83 / 2 = 87.42 years
adding in a baseline of 200 gives: 287.42 years
comparing to McGee's Utility estimate of 325 years

----

between 86048 and 101019
off by DYS391, DYS464d and DYS570
= 212.77 + 151.29 + 131.93
= 495.99 years (times two)

Or, 495.99 / 2 = 248 years
adding in a baseline of 200 gives: 448 years
comparing to McGee's Utility estimate of 500 years
