Theoretical Distribution of Genetic Distances
A Quick & Easy Method to Date STR Haplotypes

-- Sidney Sachs and William E. Howard III –

Abstract:

Several approaches have been used to estimate time relationships among Y-DNA haplotypes. The methodology of some researchers are often hard to follow and difficult to confirm and replicate. In this paper we present a straightforward methodology, based on individual DYS mutation rates, tested by models and based on Poisson statistics, in which carefully selected STR haplotype strings can be used to derive time estimates for the progenitor of the haplotype set in a process that will take only minutes to produce.

Introduction:

The traditional way to date the DNA of the Y-chromosome is to analyze a series of short tandem repeats (STR) called markers. Each marker and its mutation rate is independent of the others in the haplotype string and each is assumed to mutate randomly at its own particular rate. In the absence of evidence to the contrary, each mutation will normally change the value of the STR marker site by one unit, either upward or downward each time a mutation occurs. It is rare, but statistically predictable, to have a change of two or more units, and the average times when such larger mutations occur are predictable by Poisson statistics.

Testing agencies (e.g., Family Tree DNA) report individual testee haplotype results on their web sites with the testees' permission. Usually the haplotypes are grouped into families or clusters according to the closeness of their marker values to each other.

Determining the Theoretical Distribution of Genetic Distances:

Test results show the number of STR repeats for each marker in the haplotype. We can compare two or more haplotypes to see how close they are to each other by counting the total absolute marker differences among their results. The absolute difference between the marker values of any pair of test results is called their genetic distance (GD). Exact matches would have a genetic distance of zero. Since marker values change with time, we expect the genetic distance will increase with time. Thus the marker distance is an indicator of the time to the most recent common ancestor (MRCA) that one testee shares with another testee.

The most recent common ancestor (MRCA) of a group of testees is the most recent individual back in time from whom all males in the group have descended. An estimate can be made if the time when that MRCA lived (TMRCA) by comparing the average of all genetic distances from the average values of the group with the average of a theoretical distribution of genetic distances for different number of transitions where the number of transitions is the total number of generations that have taken place down from

the MRCA to each testee. Thus, the number of transitions involved between testees will be sum of the numbers of generations that have taken place from the MRCA down each line to each testee. Consequently, on average the number of transitions between any two testees will be twice the number of generations that have taken place. This difference between the number of generations and the number of transitions is important as we investigate the time to the MRCA.

Our method of analysis takes advantage of the randomness with which mutations take place along each marker string. In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the expectation that a given number of mutations will occur within a fixed interval of time when we know the average mutation rate, and that each mutation is independent of the past history of mutations.

Chandler (2006) derived individual marker rates for the first 37 markers that are reported by FTDNA. He also determined the average mutation rate over all 37 markers. These results cannot be used directly because when a marker changes value, it can mutate in one direction and then mutate in the opposite direction. In the absence of evidence to the contrary, each Y-DNA marker has this property of randomness. Unless we correct for this effect, we will obtain too low an estimate of the time to the MRCA of a pair or a group of haplotypes and that error will grow as we go further back in time.
Therefore, we can use Poisson statistics to derive a distribution of the number of mutations that take place within a given interval of time. Our particular challenge is at least fourfold:

- accounting for markers that randomly mutate at different rates;
- accounting for marker values that can randomly mutate either up or down;
- accounting for different TMRCAs that are derived from different lengths of haplotypes;
- accounting for computational complexities that occur when the number of mutations become very large.

We use a model with the following characteristics: We can use any number of haplotypes in the analysis. Larger numbers lead to smaller uncertainties in the parameters. We limit the theoretical calculations to 500 transitions to reduce complexity. Although this method can be employed to analyze any haplotype length, we use haplotype lengths of 37 markers in this paper since the Chandler rates are probably better known than for longer haplotype strings. The main two features of the model are (1) we use the Chandler mutation rate for each DYS site, and (2) we build in the feature that mutations can go either up or down when they occur. The model can account for different ratios of up-to--down mutations, but in the absence of evidence to the contrary, we set the ratio to unity.

Our model steps one mutation at a time, starting with zero mutations. To take into account that the mutations can be up or down, the probability for the next mutation value will be one-half upwards and one-half downwards. Knowing that a given number of mutations have occurred, we can compute the genetic distance distribution for that

number of mutations.

Table 1 was developed in the above way, one mutation at a time starting with zero mutation. Except for when the previous generic distance value is zero, the new value for the genetic distance is one half the probability of the sum of one less and one more from the one lower number of mutations. However, if genetic distances are zero, then all of probability from zero is used as part of genetic distance of 1. Table 1 shows the genetic distance listed in rows down the table for a fixed number of mutations in columns across the top of the table.

Table 1: Distribution of Genetic Distance for any Given Marker for a Fixed Number of Mutations. This distribution can be continued for larger numbers of mutations and for larger values of Genetic Distance.

No. of Mutations ⇨ / Genetic Distance ⇩	0	1	2	3	4	5	6	7	8
0	1	0	0.5	0	0.375	0	0.3125	0	0.2734375
1	0	1.0	0	0.75	0	0.6250	0	0.546875	0
2	0	0	0.5	0	0.500	0	0.46875	0	0.4375000
3	0	0	0	0.25	0	0.3125	0	0.328125	0
4	0	0	0	0	0.125	0	0.1875	0	0.2187500
5	0	0	0	0	0	0.0625	0	0.109375	0
6	0	0	0	0	0	0	0.0313	0	0.0625000
7	0	0	0	0	0	0	0	0.015625	0
8	0	0	0	0	0	0	0	0	0.0078125

The Poisson distribution for a given number of events "x" occurring in a fixed interval of time with a known average rate "a" is given by the formula:

$$\mathbf{F(x) = (a^x e^{-a})/x!}$$

where a is the mutation rate, x is the number of mutations and e is the base of the natural logarithm, 2.71828. This function can be programmed into a computer spreadsheet. John Chandler (2006) determined individual rates of mutations for each of the first 37 markers used by Family Tree DNA by studying how they changed when mutations occurred between many pairs of fathers and sons. In our approach, for a given number of transitions, the Chandler rates only need to be multiplied by that number of transitions to get the mutation rate “a” for that number of generations. This process produces the distribution for specific numbers of mutations.

After getting the distribution of mutations by using the Poisson formula, the next step is to get the distribution of genetic distance. We do this by summing across each value in the rows of Table 1 after multiplying those different values by the corresponding probability distribution of mutations. Then the final step to get the mean genetic distance

for this one marker is to get the sum each genetic distance probability by multiply by its number of mutations. Since haplotypes are defined as a set of given markers, and since we have the mean genetic distance for each of the markers separately, to get the theoretical mean distance for the haplotypes we need only to add together the individual means of these markers. This process can be done in a spreadsheet. Figure 2 shows the mean genetic distances for the 37 FTDNA markers at different numbers of transitions. A different set or number of markers would produce a different graph.

Since we have the probability of each value of genetic distances for each marker in the haplotype for a given number of generations, we also have the ability to compute the probability of getting different values of the haplotype's genetic distances. For its genetic distance of zero, this is obtained by multiplying by the probability of all markers' exact matched values. For genetic distance of 1, one of the markers most have a value of one with the rest of the markers being zero. The easiest way to get this value is to multiply the genetic distance of the zero value by the sum of adding of the probability distance of 1 for each marker divided by the percentage of zero for the same marker. For haplotypes with a genetic distance of two, there are two ways we can get this value. The first way is when one of the markers shows a genetic distance of two; the second way is when two markers each have a genetic distance of one while all the other markers are zero. The authors have done this in the same spreadsheet as above but results for genetic distances of two are not included in this paper. For genetic distances of three, there are three separate ways, and the computing requirements goes go exponentially. For this reason the authors decided to use only the mean genetic distance when dating a sample of haplotypes.

The Dating Method:

To estimate the time when the most recent common ancestor of a group of males lived, we use the following procedure:

- List the ~~37-~~marker haplotype for each testee along separate rows in a spreadsheet. Each column will show the marker value at a particular STR site. This produces a matrix of markers where each row shows an individual testee marker string and each column shows marker values at the same STR site.
- Derive the average value for each vertical column (i.e., for each STR site).
- Derive the absolute difference between each marker entry and its average value. This will result in another matrix of absolute marker differences from their respective average values.
- Sum the absolute values of the marker differences horizontally along each haplotype row, keeping fractional values. This sum will place a column to the right of the last of the 37 marker differences.
- Compute the average of the vertical listing of sums in the column formed in the last step. This is the average observed genetic distance.
- If you are using the same 37 markers that Chandler used, from Figure 2, find the corresponding number of transitions from the progenitor down to the time of the testee. This will be the average number of generations from the MRCA to the

time when the average testee was tested.

- The estimated time to the group's MRCA will be the average number of generations found in the last step times the average number of years per generation.

In setting up the input prior to analysis, the following issues and potential limitations must be considered:

- The distribution of genetic distances was based on only single step mutations from a random sample. We know that there may be mutations in our sample that are not single, but several steps.
- The multi-marker DYS 464 may have different numbers of copies.
- Some testing agency reports may contain STR sites with a fractional value, such as 14.2, rather than a whole integer. In this case, dropping the fractional value may cause some mutations to have been missed.
- The inclusion of zero values in the haplotype string will affect the computed mean of a group.
- Some companies may report some of the markers differently. For example, Family Tree DNA uses for the marker DYS 395ii, the sum of DYS 395i and another copy of it, while other companies may report only the value of the other copy.
- The sample we select to analyze may contain various degrees of bias or non-randomness. If the information comes from a project in which known relationships already are known to exist among the testees, the means we derive will be lower than if the selection process were completely random.
- The sample may contain the effects of one or more genetic bottlenecks that have occurred from the progenitor to the present time.

The authors realize that no fixed method will address these limitations since there are many varieties of both samples and analytic goals. In some cases, a researcher may elect to delete some of the samples or some markers; others may elect to modify a marker's genetic distance. To minimize the effects of these limitations we make the following comments and suggestions:

Issue 1: Use the procedure described here that takes the existence of multiple steps into account.
Issue 2: Take only the four DYS 464 values in the order reported by FTDNA or drop all DYS 464 markers, generating a new distribution of only 33 markers.
Issue 3: Choose Excel's average calculation rather than use the modal value. If the average value is used instead of the mode, the results will show smaller standard deviations.
Issue 4: Either discard the haplotypes with zero values, or change them to represent the average value of the sample, depending on the goals of the analysis, which will depend on the importance of including testees with zero values. If zero values are valid, they may be used to separate groups of haplotypes that may have significance to the study.
Issue 5: Use the FTDNA sequence or use the marker value of DYS 395ii less the

marker value of DYS 395i.
Issue 6: Depending on the goals of the analysis, retain or eliminate duplicate haplotypes in the sample.
Issue 7: Genetic bottlenecks will usually become apparent only if the sample is large. Identifying bottlenecks may be a goal of the analysis; it may be the result of a serendipitous discovery -- an unintended product of the analysis.

Whatever decisions the researcher has made, it is important to describe the details of both the setup and the analysis.

Figure 2: The Relation, Derived from the Model, Between the Average Observed Genetic Distance and The Number of Transitions for the Set of FTDNA's 37 Markers. (The average number of transitions or generations from the most recent common ancestor to the time when the average testee in the group was tested (~1945 CE)). Appendix A gives this figure on a log-log scale with reversed coordinates. Appendix B lists the data from which the figures were derived.

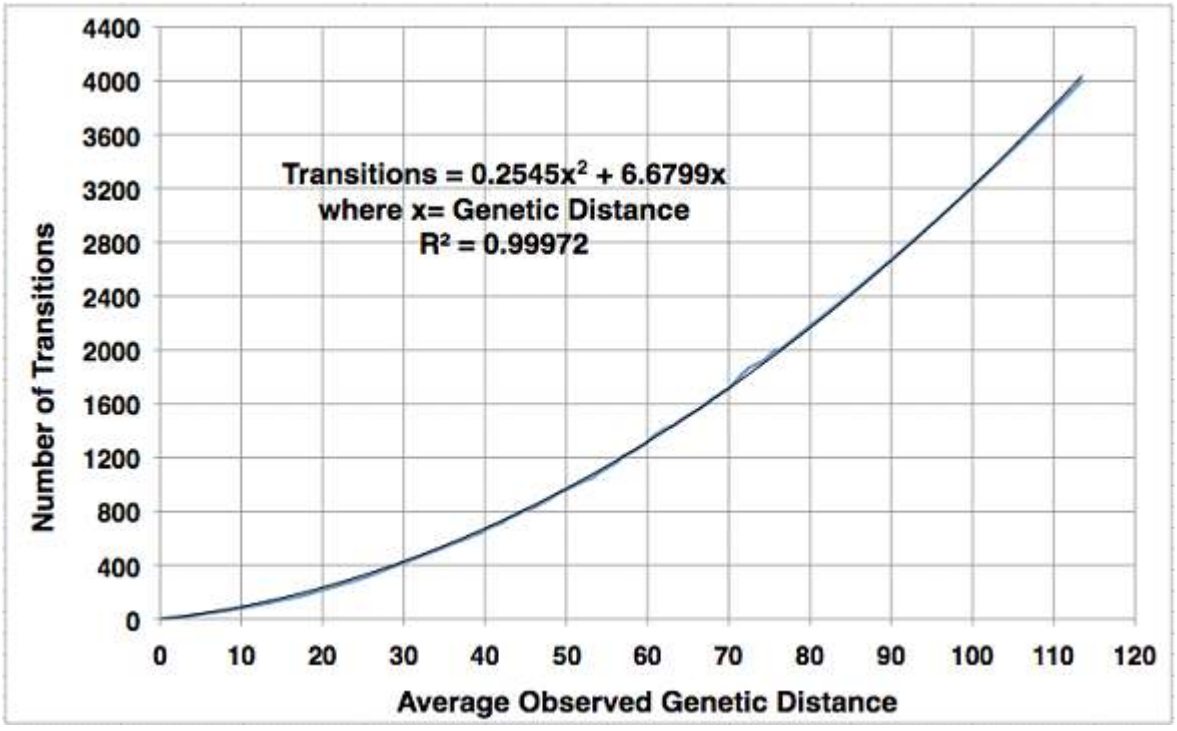
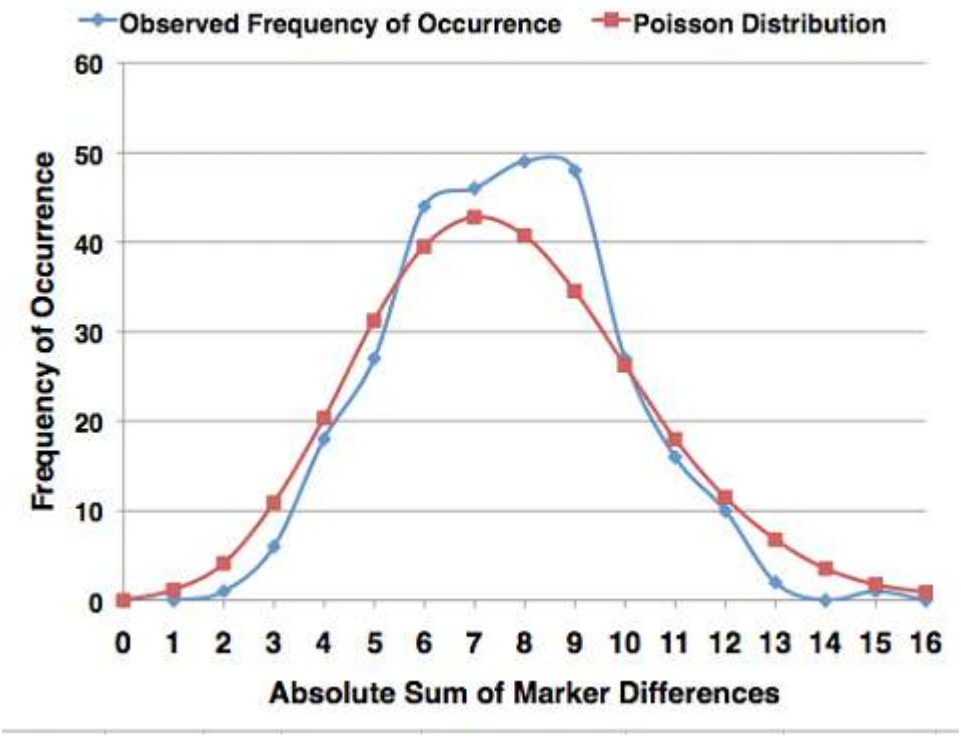


Figure 3: A Histogram of the Absolute Sum of 37 Marker Haplotype Differences of 295 SNP-tested Haplotypes of the M222 SNP.



Discussion:

Both authors have undertaken the challenge to derive a genealogical time scale that extends into times of interest to geneticists -- one, (SS) using the forgoing methology and one (WEH) using the RCC correlation approach (Howard 2009 and subsequent papers). This paper incorporates mutation rates of the first set of 37 markers reported by FTDNA. Any haplotype marker length can be used if we know each marker's mutation rate. We compared the methodology reported here with the methodology of the RCC correlation approach by using the same large set of 295 M222 haplotypes, all of which have been SNP-tested, and determining their TMRCAs separately. The date estimates using the approach in this paper was derived from an average observed mean genetic distance of 7.6041 (standard deviation 1.7%), which corresponds to 65.5 transitions. Figure 3 shows that the observed distribution of the absolute sum of marker differences is in good agreement with the distribution predicted by a Poisson distribution. Since, on average, the number of transitions between testees is expected to be twice the number of generations that have taken place, an average of 65.5 transitions is equivalent to 38 generations. If one generation is 28.8 years, the progenitor of the M222 haplotypes selected here is estimated to have lived about 1100 years ago. In another paper (Howard and McLaughlin 2011) the date of origin of the M222 SNP, derived from 320 and 683 database samples was RCC ~ 84 or 3640 years ago with an estimated standard deviation (SD) of 300 years (18%). The smaller comparison sample of 295 haplotypes led to an RCC value of 87.5 or 3790 years ago.

Acknowledgements:

We wish to thank J. J. (Jim) Logan for very helpful comments and criticisms of this paper and to thank Fred Schwab for his contributions that model the effects that the behavior of mutations have with the times we derive.

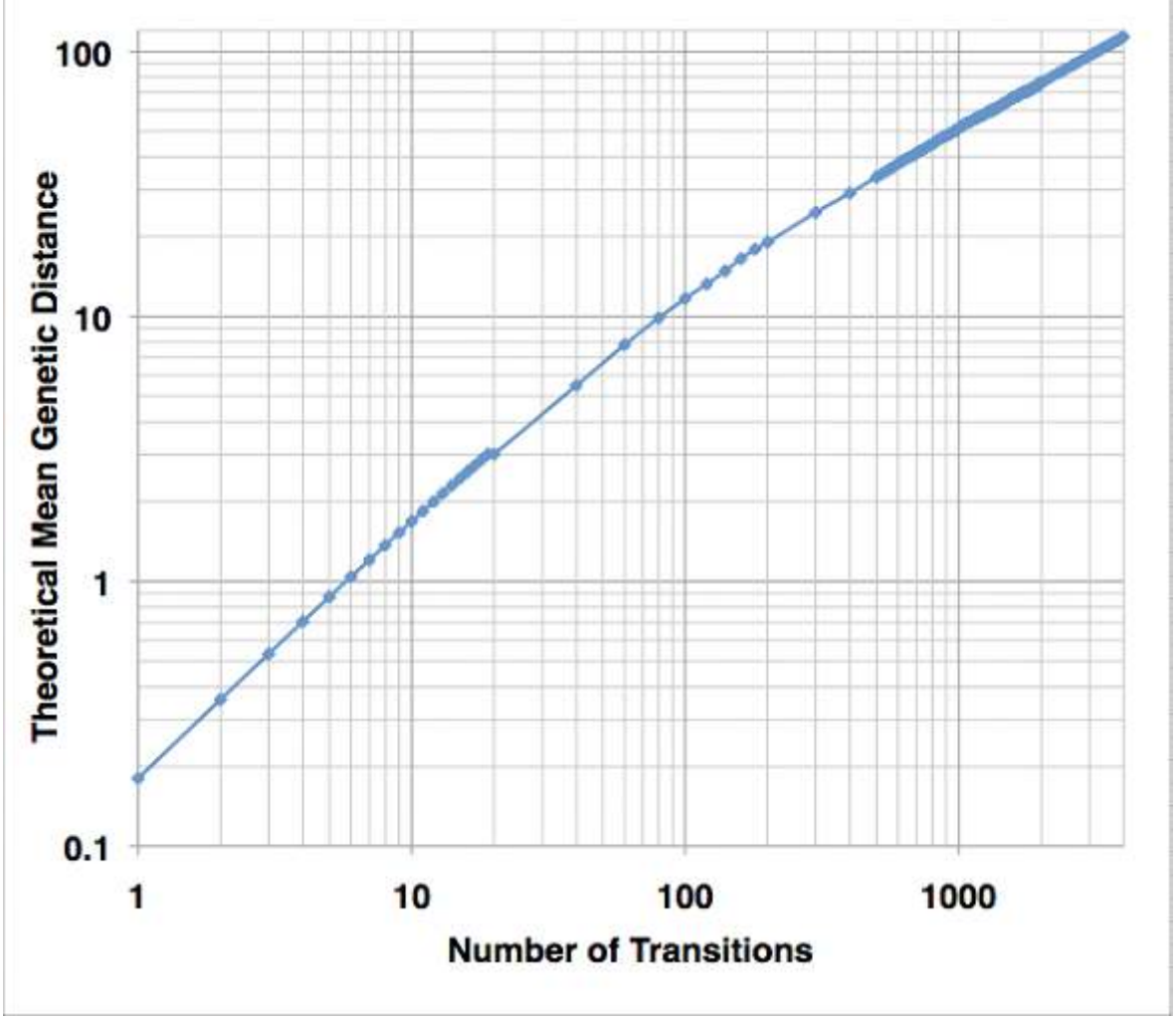
References:

Howard, William E. III, The Use of Correlation Techniques for the Analysis of Pairs of Y-STR Haplotypes, Part 1: Rationale, Methodology and Genealogy Time Scale, Journal of Genetic Genealogy, 5(2)256-270, 2009: http://www.jogg.info/52/files/Howard1.pdf

Howard, William E. III and McLaughlin, John D., A Dated Phylogenetic Tree of M222 SNP Haplotypes: Exploring the DNA of Irish and Scottish Surnames and Possible Ties to Niall and the Uí Néill Kindred, Familia, Ulster Genealogical Review No.27 (2011), pp. 14-50. Ulster Genealogical & Historical Guild.

Appendices:

Appendix A:
This figure is a log-log plot of Figure 2 with coordinates exchanged. Note that the entries consist of a combination of model and theoretical results, which were found to fit well at their junction points. Researchers may find this chart easier to use than Figure 2.



Appendix B:

Column 2 gives the average observed genetic distance for the number of transitions listed in Column 1. Column 3 gives the fraction of time for each transition that you will observe that no markers have changed. Column 4 gives the fraction of time that you will observe that one marker has changed. These results apply for the set of the first 37 FTDNA marker sequence.

No. of Transitions Change	Mean GD	Genetic Distance No Marker Change	Genetic Distance One Marker
1	0.18055	0.834182	0.151878
2	0.358128	0.696941	0.253734
3	0.532861	0.583182	0.318376
4	0.70487	0.488747	0.355605
5	0.874269	0.410237	0.372892
6	1.041164	0.344869	0.375909

7	1.205658	0.290361	0.368944
8	1.367847	0.244841	0.355217
9	1.527822	0.20677	0.337128
10	1.685669	0.174883	0.316449
11	1.841471	0.148134	0.294474
12	1.995304	0.125663	0.272134
13	2.147243	0.106758	0.250083
14	2.297356	0.090829	0.228768
15	2.445711	0.077389	0.208481
16	2.592369	0.066032	0.1894
17	2.73739	0.056421	0.171619
18	2.88083	0.048277	0.155172
19	3.022744	0.041365	0.140048
20	3.163182	0.035491	0.12621
21	3.302193	0.030492	0.113597
22	3.439823	0.026233	0.102141
23	3.576116	0.022597	0.091763
24	3.711122	0.019491	0.082384
25	3.844853	0.016833	0.073924
26	3.977376	0.014556	0.066305
27	4.108718	0.012602	0.059452
28	4.238911	0.010924	0.053296
29	4.36799	0.009481	0.047771
30	4.495985	0.008237	0.042816
35	5.120708	0.004149	0.024783
40	5.722303	0.002145	0.014416
45	6.3034	0.001136	0.00845
50	6.866138	0.000615	0.004999
55	7.412276		
60	7.943278		
65	8.460375		
70	8.964619		
75	9.456915		
80	9.938052		
85	10.40872		
90	10.86954		
95	11.32106		
100	11.76377		
120	13.45495		
140	15.03507		
160	16.5212		
180	17.9263		
200	19.2603		
250	22.33118		
300	25.08667		
350	27.5801		

400	29.83763
450	31.84643
500	33.55954
600	37.82
700	41.34
800	44.61333
900	48.09333
1000	51.3
1200	56.78
1400	61.76667
1600	67.23333
1800	71.38
2000	76.18970967
2500	86.6260774
3000	96.20588078
3500	105.1274446
4000	113.5216971

Note to Appendix B:

A separate computer model (derived by Sachs) that presented results out to 4000 transitions produced results that were almost identical to the results for the separate model that ended with 500 generations. Thus we can probably safely expand the 500-generation model out to 4000 transitions by equating the results of both models to the composite result presented here.