

---

Journal: [www.joqq.info](http://www.joqq.info)

Originally Published: Volume 2, Issue 2 (Fall 2006)

Reference Number: 22.005

# HAPLOGROUP PREDICTION FROM Y-STR VALUES USING A BAYESIAN-ALLELE- FREQUENCY APPROACH

Author(s): T. Whit Athey

# Haplogroup Prediction from Y-STR Values Using a Bayesian-Allele-Frequency Approach

T. Whit Athey

A new Bayesian allele-frequency approach to predicting the Y-chromosome haplogroup from a set of Y-STR marker values is presented and compared to other approaches. The method has been implemented in an Excel-based program, where an arbitrary number of STR markers may be input and a “goodness of fit” score for 15 haplogroups (C, E3a, E3b, G, H, I1a, I1b1, I1b2, J, L, N, O, Q, R1a, and R1b) and the Bayesian probability for each haplogroup is returned. This method has been applied to 100 R1b haplotypes, 50 I1a haplotypes (all having 37 STR markers available), and 54 of the YCC sample set (having 20 STR markers available).

Version: 20 October 2006

## Introduction

There is considerable interest in determining the Y-chromosome *haplogroup*, a group or family of Y-chromosomes related by descent and defined by the pattern of *single nucleotide polymorphisms* (SNPs), from the Y-STR haplotype. Two methods have been described previously, an allele-frequency-goodness-of-fit approach and a genetic-distance approach (Athey 2005).

While the allele-frequency-goodness-of-fit has been fairly successful in indicating the haplogroup, it does not actually provide a prediction or probability that a haplotype is in a particular haplogroup. The present article describes a Bayesian approach based upon allele frequencies. Such an approach does result in the probability that a Y-STR haplotype is in a haplogroup.

## Methods

### Allele Frequencies

The allele frequencies required for the present approach were calculated from collections of haplotypes extracted from published articles and public databases as described by Athey (2005). Haplogroup prevalence information was also obtained from these sources.

### Bayes Theorem Generalized

Suppose that a certain outcome,  $S$ , which normally occurs in the general population with probability  $P(S)$ , is correlated with the result of a particular test  $B$ . Bayes' theorem tells us that:

$$\Pr(S|B) = \frac{\Pr(B|S)\Pr(S)}{\Pr(B|S)\Pr(S) + \Pr(B|\text{NOT } S)\Pr(\text{NOT } S)} \quad 1$$

In the above expression,  $\Pr(S|B)$  is to be read as “probability of outcome  $S$ , given that we have obtained test result  $B$ .” Similarly,  $\Pr(B|S)$  is read as “probability of test result  $B$ , given that we know that the state is  $S$ .”

We now must generalize Bayes' Theorem for multiple possible outcomes, multiple possible test results, and more than one test. The multiple outcomes will correspond to multiple possible haplogroups, the multiple test results will correspond to the different possible values of a Y-STR test on a single marker, and the multiple tests will correspond to the many different Y-STR marker tests that are available.

The Wikipedia article on Bayes' Theorem gives the generalization to the case where a set  $\{S_i\}$  forms a partition of the event or outcome space (that is, there are more than two possible outcomes or states):

$$\Pr(S_i|B) = \frac{\Pr(B|S_i)\Pr(S_i)}{\sum_j \Pr(B|S_j)\Pr(S_j)} \quad (2)$$

for any  $S_i$  in the partition (for any of the possible states,  $S_i$ ).

Note that the test  $B$  may have more than two possible results,  $B_j$ , and there will be an expression like the above for each of the possible results  $B_j$ .

If there are two tests,  $T_1$  and  $T_2$ , that may have predictive value for the state  $S_i$ , we can consider that the probability of state  $S_i$ , given test results  $T_1$  and  $T_2$ , or  $\Pr(S_i|T_1 \text{ AND } T_2)$ , can be written:

---

Address for correspondence: [wathey@hprg.com](mailto:wathey@hprg.com)

$$\Pr(S_i | T_1 \text{ AND } T_2) = \frac{\Pr(T_1 \text{ AND } T_2 | S_i) \Pr(S_i)}{\Pr(T_1 \text{ AND } T_2 | S_i) \Pr(S_i) + \Pr(T_1 \text{ AND } T_2 | \text{NOT } S_i) \Pr(\text{NOT } S_i)} \quad (3)$$

If we were considering the entire population, instead of a sample of the population, the numerator would represent the total number of haplotypes in the  $S_i$  haplogroup that match the test haplotype, and the denominator would represent the total number of haplotypes of any haplogroup that match the test haplotype.

If the tests are independent, the right side of the equation may be simplified. Tests of different Y-STR markers would generally appear to satisfy the independence condition, but there are important exceptions discussed below. The right side of the equation becomes for independent tests:

$$\frac{\Pr(T_1 | S_i) \times \Pr(T_2 | S_i) \Pr(S_i)}{\Pr(T_1 | S_i) \Pr(T_2 | S_i) \Pr(S_i) + \Pr(T_1 | \text{NOT } S_i) \Pr(T_2 | \text{NOT } S_i) \Pr(\text{NOT } S_i)}$$

Or, even more generally, for any number of independent tests and any number of outcome states, our generalized version of Bayes' Theorem becomes:

$$\Pr(S_i | T_1 \text{ AND } T_2 \text{ AND } T_3 \text{ AND } \dots \text{ AND } T_n) = \frac{\Pr(T_1 | S_i) \Pr(T_2 | S_i) \Pr(T_3 | S_i) \times \dots \times \Pr(T_n | S_i) \Pr(S_i)}{\sum_j [\Pr(T_1 | S_j) \Pr(T_2 | S_j) \Pr(T_3 | S_j) \times \dots \times \Pr(T_n | S_j) \Pr(S_j)]} \quad (4)$$

Let the  $T_k$  represent tests of the  $k$ th Y-STR marker, each of which will return an integer result (out of several possible integer results). Then we wish to find the probability that the haplotype (complete set of Y-STR values) is from Haplogroup  $S_i$ . Following the above, we can identify the quantities in the last expression above as follows:

$T_k$  represents the test result for the  $k^{\text{th}}$  Y-STR marker (for example,  $T_1$  might represent  $\text{DYS393} = 13$ , and  $T_2$  might represent  $\text{DYS390} = 23$ ).

$\Pr(S_i | T_1 \text{ AND } T_2 \text{ AND } T_3 \text{ AND } \dots \text{ AND } T_n)$  is the probability that the unknown haplotype is in Haplogroup  $S_i$ , given that we know the test results  $T_1, T_2, T_3$ , etc.

$\Pr(T_1 | S_i)$  is the probability of the result  $T_1$  occurring (e.g.,  $\text{DYS393} = 13$ ) in Haplogroup  $S_i$ . This will be equal to the allele frequency (in that haplogroup) for that allele value.

$\Pr(S_i)$  is the probability (prior to any test results) of the person being from a particular haplogroup. If we don't have any test results, then our best estimate of the probability of a particular haplogroup is just its frequency in the general population from which the person comes. Note that this may be quite different in different parts of the world.

### Illustrative Examples

The Bayesian approach outlined above will now be applied in detail to show how it works. We will consider an artificial case where we have just two Y-STR marker values and we will calculate the probability that the "haplotype" is in one of four haplogroups, G, J, I1a, or R1b.

**Tables 1, 2, and 3** contain the data that we need to collect before Equation 4 can be applied. Basically, we need to know the frequency in the population of the four haplogroups and the allele frequencies for the two markers for each of the haplogroups. We will assume that the person being tested and whose haplogroup will be predicted, is a person from northwest Europe or the United States, and the frequencies of each of these haplogroups will be chosen to be approximately those actually observed, except that the four frequencies have been scaled so that they add to 1.00. That is, we will assume that the only possibilities are the four named haplogroups and that the population only contains those four haplogroups. **Table 1** shows these assumed values, which would be approximately the actual frequencies in a northwest European population (if those were the only haplogroups).

**Table 1 Assumed Population Frequencies**

	Haplo-group G	Haplo-group J	Haplo-group I1a	Haplo-group R1b
Pop. Freq	0.02	0.04	0.15	0.79

**Table 2 Allele Frequencies for DYS393**

Repeat Values	Haplo-group G	Haplo-group J	Haplo-group I1a	Haplo-group R1b
11	0.003	0.007	0.001	0.000
12	0.013	0.884	0.022	0.021
13	0.204	0.092	0.876	0.950
14	0.680	0.015	0.088	0.028
15	0.095	0.002	0.012	0.001
16	0.005	0.000	0.001	0.000

**Table 3 Allele Frequencies for DYS390**

Repeat Values	Haplo-group G	Haplo-group J	Haplo-group I1a	Haplo-group R1b
20	0.010	0.001	0.001	0.000
21	0.116	0.008	0.009	0.001
22	0.603	0.126	0.610	0.010
23	0.254	0.563	0.345	0.279
24	0.013	0.231	0.033	0.561
25	0.004	0.062	0.002	0.142
26	0.000	0.009	0.000	0.007

Let's first avoid the use of the formulas and calculate from basic principles the probability of each haplogroup, assuming that we only know the value on DYS393. This is a very artificial situation, but it illustrates some important points. Suppose that our testee has a value of 14 on DYS393. We may be tempted at first glance to say that Haplogroup G is most likely. After all, 14 is the modal value in G for DYS393. However, let's see how this plays out.

Assume that we pick 1000 people at random from a population that has only the four haplogroups represented, each of which has the frequencies shown in **Table 1**. On average, such a sample of 1000 people would have 20 persons in G, 40 in J, 150 in I1a, and 790 in R1b.

Of the 20 people in G, on average (using **Table 2**), about 14 of them would have a value of 14. Of the 40 people in J, about 1 would have a value of 14. Of the 150 people in I1a, about 13 of them would have a value of 14. Finally, of the 790 people in R1b, about 22 would have a value of 14.

So, we have the possibly surprising result, that of the people with 14 on DYS393 in our sample of 1000, the largest number would actually be in R1b, not in G, in spite of the fact that DYS393=14 is an unusual value in Haplogroup R1b. In our sample of 1000, we would have a total of 50 from all haplogroups with DYS393=14, so we would have the following results, using the notation of our earlier development:

$$\Pr(G \mid \text{DYS393}=14) = 14/50 = 28\%$$

$$\Pr(J \mid \text{DYS393}=14) = 1/50 = 2\%$$

$$\Pr(I1a \mid \text{DYS393}=14) = 13/50 = 26\%$$

$$\Pr(R1b \mid \text{DYS393}=14) = 22/50 = 44\%$$

However, we are not restricted to just one marker to make our predictions, and it is the availability of multiple markers that makes haplogroup prediction possible from Y-STR markers.

Before we go on to consider more markers, let's check to see that the formulas that we developed give the same probabilities that we obtained above for just the one marker.

The formula for the probability of Haplogroup G would be:

$$\Pr(G \mid \text{DYS393}=14) = P(14 \mid G)\Pr(G) / D$$

where

$$D = P(14 \mid G)\Pr(G) + P(14 \mid J)\Pr(J) + P(14 \mid I1a)\Pr(I1a) + P(14 \mid R1b)\Pr(R1b)$$

Similarly,

$$\Pr(J \mid \text{DYS393}=14) = P(14 \mid J)\Pr(J) / D$$

$$\Pr(I1a \mid \text{DYS393}=14) = P(14 \mid I1a)\Pr(I1a) / D$$

$$\Pr(R1b \mid \text{DYS393}=14) = P(14 \mid R1b)\Pr(R1b) / D$$

From the tables, we see that:

$$P(14 \mid G)\Pr(G) = (.68)(.02) = .0136$$

$$- P(14 \mid J)\Pr(J) = (.015)(.04) = .00060$$

$$P(14 \mid I1a)\Pr(I1a) = (.088)(.15) = .01320$$

$$P(14 \mid R1b)\Pr(R1b) = (.028)(.79) = .02212$$

and the denominator, D, is just the total of those four quantities:

$$D = .01360 + .00060 + .01320 + .02212$$

$$= .04952$$

Substituting, we get

$$\Pr(G \mid \text{DYS393}=14) = .0136/.04952 = .275$$

$$\Pr(J \mid \text{DYS393}=14) = .0006/.05928 = .012$$

$$\Pr(I1a \mid \text{DYS393}=14) = .01320/.04952 = .267$$

$$\Pr(R1b \mid \text{DYS393}=14) = .02212/.04952 = .447$$

We see that these are the same probabilities that we calculated manually by considering the 1000 people.

Now we can calculate the probability of being in each haplogroup, given that both DYS393=14 and DYS390=22 have been found as a result of testing. Again, these are the modal values of Haplogroup G, but let's calculate the probabilities using Equation 4:

$$\begin{aligned} \Pr(G \mid \text{DYS393}=14 \text{ AND } \text{DYS390}=22) &= \\ &= \frac{\Pr(\text{DYS393}=14 \mid G) \Pr(\text{DYS390}=22 \mid G) \Pr(G)}{\sum_j [\Pr(T_1 \mid S_j) \Pr(T_2 \mid S_j) \Pr(S_j)]} \\ &= (.68)(.603)(.02) / D \end{aligned}$$

where

$$\begin{aligned} D &= (.68)(.603)(.02) + (.015)(.126)(.04) + \\ &\quad + (.088)(.610)(.15) + (.028)(.010)(.79) \\ &= .00820 + .00007 + .00805 + .00022 = .01654 \end{aligned}$$

So,

$$\begin{aligned} \Pr(G \mid \text{DYS393}=14 \text{ AND } \text{DYS390}=22) &= \\ &= .00820 / .01654 = 0.496 \end{aligned}$$

Similarly,

$$\Pr(J \mid \text{DYS393}=14 \text{ AND } \text{DYS390}=22) = 0.004$$

$$\Pr(I1a \mid \text{DYS393}=14 \text{ AND } \text{DYS390}=22) = 0.487$$

$$\Pr(R1b \mid \text{DYS393}=14 \text{ AND } \text{DYS390}=22) = 0.013$$

Bringing the second marker value into consideration dramatically reduces the probability of R1b, while boosting it for G and I1a. Adding more markers would further refine the probabilities and provide discrimination between I1a and G.

With the earlier approach to haplogroup prediction, which calculated a 'goodness of fit' score for the haplotype for each haplogroup (Athey, 2005), adding more markers did not always result a higher score for the highest scoring haplogroup. After a couple of dozen markers, the "fitness" score typically remained about the same because the fitness is averaged over all markers. However, with the Bayesian approach, more markers will usually improve the probability (but only if the added markers actually provide discriminative power). Typically, only 10-20 markers are sufficient to bring the probability for one of the haplogroups up to a value in excess of 99%. In contrast, the fitness score sometimes was almost the same for the two highest-scoring haplogroups, for example, in cases where R1b and Q both show moderate scores.

### Independence of Y-STR Markers?

In the development above, the assumption was made that the test results for Y-STR markers were independent. This does not mean that particular values should not be characteristic of particular haplogroups—the whole approach depends on that fact. The independence assumption means that there is no correlation of marker values *within* one of the haplogroups included in the analysis. The independence assumption greatly simplifies the development and the calculation of probabilities—indeed, it makes the whole approach feasible. However, the fact is that the markers are not always independent within haplogroups. There are a number of cases where there are "varieties" of haplotypes within a haplogroup, and these varieties are usually characterized by some correlation of values at two or more markers. Founder effects and population dynamics cause, for example, DYS390 and DYS462 to be highly correlated within Haplogroup I1a. When this happens, the assumption that

$$\Pr(\text{DYS390}=22 \text{ AND } \text{DYS462}=12 \mid I1a) =$$

$$\Pr(\text{DYS390}=22 \mid I1a) \Pr(\text{DYS462}=12 \mid I1a)$$

is no longer true. Using the available data on allele frequencies, the left side of the expression above is about 0.71, while the right side is equal to about 0.43, a difference of almost a factor of two. The difference is even more pronounced for the low probability combinations of values.

Does this non-independence of marker values cause the overall approach to fail? The answer is, sometimes yes and usually no. The answer is yes if a small number of markers is being analyzed, if two of the markers are very correlated, and if it is important to obtain an accurate value for the probability (i.e., rather than simply, a result, for example, "greater than 95%").

The answer is no if we have test results for a large number of markers, because after 15-20 markers have been added to the analysis, the probability for one haplogroup will usually be over 99% regardless of any correlated markers or unusual values. If we get a result of 99%, we usually don't care if it is 99.0% or 99.99%. So, one practical solution to the problem of non-independence of markers is to add markers to the analysis until the probability for some haplogroup has been "driven" well past 99%.

There is another approach to the non-independence problem. If each variety or subhaplogroup that has non-independent markers is treated as a separate haplogroup in the analysis, then the independence assumption again is a good one. If adequate data on each variety is available, then this is a reasonable

approach. However, for some haplogroups and subclades, it is difficult to obtain the necessary data. A subclade version of the program is planned for the future.

### Other Practical Problems

Any allele-frequency approach depends upon having available the allele-frequency distributions for each marker in each haplogroup. For the major haplogroups, there is an abundance of data. For the minor and non-European haplogroups, the data available is scarcely sufficient for the haplogroup to be included, especially for the markers that are not often measured.

## RESULTS

### *Results From Testing R1b Haplotypes*

The Bayesian algorithm, with 15 haplogroups and their associated allele frequency distributions considered, was applied to 100 haplotypes of 37 markers each from Y-Search where the haplogroup had been indicated as R1b. This is the same R1b dataset that was used in Athey (2005), less the one haplotype that was determined to be not in R1b in that study.

The scores from the haplogroup fitness algorithm for this R1b set of haplotypes ranged from 40 to 85 in the previous study (Athey, 2005). The mean of the scores was found to be about 65.

Applying the Bayesian algorithm to the same 100 R1b haplotypes, using northwest European priors (haplogroup frequencies in the northwest European population), resulted in probabilities of R1b that were all greater than 99%.

### *Results from Testing Fifty I1a Haplotypes*

In Athey (2005), 50 haplotypes with 37 Y-STR marker values were identified on Y-Search that had a DYS455 repeat value of 7, 8, or 9 (generally considered to indicate membership in Haplogroup I1a), all with different surnames. One haplotype had a value of 9 for DYS455 and the other 49 had the value of 8.

The I1a fitness scores for the 50 haplotypes were reported in Athey (2005) to range from 31 to 89, with an average score of about 65. Only four of the haplotypes had scores less than 50, each of which, indeed, had somewhat unusual haplotypes.

When the Bayesian algorithm was applied to these same 50 haplotypes, using northwest European priors, the probability for Haplogroup I1a was greater than 99% in every case, even the four somewhat unusual haplotypes.

### *Application to the Haplotypes of the YCC Set of Y-Chromosome Samples*

The Y-Chromosome Consortium has collected several dozen blood samples from populations around the world and has performed SNP and Y-STR tests on them. The haplotypes for 25 markers for the samples have been reported by Butler (2002). The Bayesian algorithm was applied to the haplotypes that were from any of the 15 haplogroups included in the program. In the analysis for each YCC sample, the priors for all 15 haplogroups were chosen on the basis of the origin of the samples: Asia and Americas priors for Haplogroups C, H, N, O, and Q; western Europe priors for Haplogroups G, I, J, R1b; eastern Europe priors for N, R1a; and African priors for Haplogroups E3a, E3b (since these samples originated from Africa). Note that there is no sample in the YCC collection for Haplogroup L.

**Table 4** shows the results for the YCC set of haplotypes. In every case except one, the highest probability returned by the program was for the correct haplogroup. In 50 out of the 54 cases, the correct haplogroup was predicted with a probability of over 99%. Note, however, that YCC61 and YCC74 were labeled by the YCC as I1. The program returned results of I1b2 and I1b1 for those, which are correct as "I1," but it is not known if the subgroup is correct. Two cases where the program did not perform well are discussed below.

The YCC79 sample (Haplogroup G1) showed a probability of just 48% for Haplogroup G and a 52% probability for Haplogroup I1a. This is probably because the allele frequencies for Haplogroup G were calculated from a dataset that included very few G1 haplotypes.

The YCC03 haplotype is in Haplogroup Q. The Bayesian prediction algorithm gave the highest probability to this haplogroup, but that probability was just 61%. The second highest probability was for Haplogroup C with 39%. Note that Haplogroup C is a large and old haplogroup in Asia and it exhibits considerable diversity in its Y-STR values. This occasionally causes difficulty in distinguishing Haplogroup C from others. Haplogroup C should have been split into several of its subgroups and each included in the analysis, but there was scarcely enough data available for C to be included at all.

### **Limitations**

The chief limitation of any allele-frequency approach to haplogroup prediction is the availability of an adequate database of Y-STR haplotypes from which the allele frequencies can be calculated. For the common

haplogroups such as R1b and I1a, there are abundant data. Even for haplogroups that occur at frequencies of just 1-4% in Europe, such as E3a, E3b, G and J, there are adequate data.

For several of the haplogroups, there is substructure that must unfortunately be ignored. Ignoring substructure often leads to the non-independence of marker values as discussed above, or to overly broad allele frequency distributions. The ideal solution is to include these subgroups as separate haplogroups in the analysis, but adequate data are often unavailable. For haplogroups such as Haplogroups C, H, L, N, and Q, there is scarcely enough data for those haplogroups to be included whole. Because so few haplotypes for these haplogroups are publicly available, it is likely that those that are available may not be representative.

With increasing numbers of people around the world being tested, many of these limitations may soon be resolved.

## Conclusion

The allele-frequency approach to haplogroup prediction provides a powerful and robust alternative to genetic-distance approaches, whether through a “goodness-of-fit” method or a Bayesian probability approach. However, both allele-frequency approaches depend on having sufficient data from the haplogroups under consideration to enable the calculation of realistic allele-frequency values for each haplogroup.

## Electronic-Database Information

www.ysearch.org      genetic genealogy database  
 www.ybase.org      genetic genealogy database  
<http://www.hprg.com/hapest5/>  
                                  haplogroup predictor

## References

- [Athey TW \(2005\) Haplogroup Prediction using an allele-frequency approach. J Genetic Genealogy, 1:1-7.](#)
- [Butler JM, Schoske R, Vallone PM, Kline MC, Redd, AJ, Hammer MF \(2002\). A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. Foren Sci Int 129:10-24](#)
- [Wikipedia Encyclopedia \(2006\), article on Bayes' Theorem, especially the Alternative Forms of Bayes' Theorem.](#)

**Table 4 Results for the YCC Samples**

YCC No.	Haplo-group Designated by YCC	Calculated Probability of that Haplogroup Using Bayesian Approach (%)	Next Highest Probability (%), (if Pr $\geq$ 0.1%) and the Next Highest Haplogroup
YCC23	C3b	99.9 (C)	
YCC33	E3a	99.9 (E3a)	
YCC36	E3a	99.9 (E3a)	
YCC40	E3a	99.9 (E3a)	
YCC43	E3a	99.9 (E3a)	
YCC45	E3a	99.9 (E3a)	
YCC65	E3a	99.9 (E3a)	
YCC31	E3a1	99.9 (E3a)	
YCC44	E3a1	99.9 (E3a)	
YCC32	E3b	99.9 (E3b)	
YCC79	G1	48.4 (G)	51.4 (I1a)
YCC52	G2	99.9 (G)	
YCC53	G2	99.9 (G)	
YCC55	G2	99.9 (G)	
YCC80	G2a	99.9 (G)	
YCC24	G2a1	99.8 (G)	0.1 (E3a)
YCC58	H1	99.8 (H)	0.1 (J)
YCC61	I1	99.9 (I1b2)	
YCC74	I1	99.9 (I1b1)	
YCC63	I1a1	99.9 (I1a)	
YCC72	I1b1	99.9 (I1b1)	
YCC59	J	99.9 (J)	
YCC56	J2	95.3 (J)	4.7 (H)
YCC60	J2	99.9 (J)	
YCC77	N1	98.4 (N)	1.6 (C)
YCC47	N3a	99.9 (N)	
YCC48	N3a	99.9 (N)	
YCC51	N3a	99.9 (N)	
YCC49	N3a1	99.9 (N)	
YCC50	N3a1	99.9 (N)	
YCC66	O1	99.7 (O)	0.3 (C)
YCC67	O1	99.9 (O)	
YCC69	O2a	93.6 (O)	6.4 (C)
YCC68	O3c	99.9 (O)	
YCC57	O3e	99.9 (O)	
YCC78	O3e	99.9 (O)	
YCC02	Q	99.9 (Q)	
YCC03	Q	61.4 (Q)	38.6 (C)
YCC04	Q	99.9 (Q)	0.3
YCC25	Q	99.9 (Q)	
YCC12	Q3	99.9 (Q)	
YCC13	Q3	99.7 (Q)	0.2(C)
YCC15	Q3	99.7 (Q)	0.3 (C)
YCC16	Q3	99.7 (Q)	0.3 (C)
YCC17	Q3	99.9 (Q)	
YCC18	Q3	99.9 (Q)	0.1 (R1b)
YCC14	Q3c	99.9 (Q)	
YCC70	R1a	99.9 (R1a)	
YCC81	R1a	99.9 (R1a)	
YCC26	R1b	99.9 (R1b)	
YCC27	R1b	99.9 (R1b)	
YCC62	R1b	99.9 (R1b)	
YCC64	R1b	99.9 (R1b)	
YCC71	R1b	99.9 (R1b)	0.1 (Q)