

Geographic Patterns of Haplogroup R1b in the British Isles

Kevin D. Campbell

Abstract

The recent availability of Y-STR databases has provided the opportunity to further explore geographic and subclade patterns of Haplogroup R1b in the British Isles. However, works based upon this data leave gaps in the published analyses that make it difficult to link conclusions to the supporting data. This paper identifies haplotypes in the Oxford Genetic Atlas Project data and links these to Bryan Sykes' book, *Blood of the Isles*. The analysis provides support for Sykes' conclusions and posits the genetic signatures for the Dal Riada Celts and Picts identified in Sykes' book.

Introduction

While DNA testing has evolved rapidly, there is a dearth of reliable Y-STR DNA data for serious analysis. This absence is particularly surprising given the heavy concentration of the R1b haplogroup in Western Europe and the bias of certain key databases such as YSearch, SMGF, and FTDNA toward participants from North America and Western Europe.

The lack of reliable data can be attributed to a number of causes. These include: inconsistent use of markers and nomenclature, cost involved with extensive panels of markers, and a number of other issues that are familiar to most academic and amateur genetic researchers. However, it is suggested that there are two root causes that have significantly hindered population analysis – (1) the lack of uniformly collected, independently verified data sets and (2) the tendency of researchers to shield and obfuscate¹ their analysis.

The major factor contributing to the first is the nature of the submission process at the largest databases, which results in non-validated and unsubstantiated data. The sad truth is that many of this field's largest databases rely on data and geographic inputs provided by enthusiastic, but uninformed individuals. While transcription and upload errors for YSearch and

¹ In the case of academic researchers, many keep tight control of their data. It is hoped that in the future that scientific journal editors will require the submission of supporting data, which they might even hold for a period of time even after publication of analytic articles. Such submissions would ensure that important research is fully documented even if it is years later when the article is no longer at the forefront of everyone's mind.

Address for correspondence: Campbell@alum.mit.edu

Received: January 1, 2007; accepted: April 27, 2007.

SMGF may be very low, many errors creep in related to marker translation and geographic speculation. Though some might dismiss marker translation as insignificant, errors in geographic data are not. In fact the essence of population analysis is the deduction and inferences that can be made about how the Y-STR data relates to specific geographic areas or population migrations.

As an example, in the Campbell project 73% of the participants (123 out of 169) do not have a reliable paper trail that places their genetic ancestor in either Scotland or Ireland. Of those that do, 39% have a paper trail or some level of confidence that their ancestors came from Ireland and 61% believe that their ancestors came from Scotland. Given that the majority of Campbell participants can only reliably document the birth location of their oldest proven ancestor to the 18th and 19th centuries,² it is also likely that even much of the Ireland/Scotland split is further confused by the Scotch-Irish migrations of that period. In short, it is suggested that much of the location data provided by DNA participants to the large voluntary databases is, at best, of marginal reliability when performing large population genetic studies.

The aforementioned statistics call into question the reliability of information provided by individual participants and suggests that possible organized and structured data gathering studies might provide better sources of data. However, some respected researchers have published popular texts with new theories while providing insufficient information or omitting critical linkages that might facilitate a formal and critical review.

For example, though Capelli's study of Britain has been published for several years (Capelli et al, 2003), only

² For the Campbell Project, the distribution of the birth year of the oldest proven ancestor of the participants is as follows. 4% earlier than the 1600s, 7% in the 1600s, 56% in the 1700s, and 33% in the 1800s. There is no reason to expect this to be anything but representative, and in fact, one could be convinced that some of these participants are outliers with longer than usual paper genealogies.

recently has the underlying data been made available. While EthnoAncestry³ has introduced a DNA test for “Pictishness” they have not yet published any details regarding their analyses.

Though Bryan Sykes has made the data available that he used in his recent popular book “Blood of the Isles”, he also leaves out critical linkages between the particular haplotypes and conclusions that he draws in the text. This is particularly frustrating since while the haplogroups in his study are reasonably distinguishable, the assertions and theories that he makes about specific subclades are not easily reviewed in the context of the supporting data. The present article will fill in some of this missing information.

Similarly, most people would consider the recent book, *The Origins of the British*, by Stephen Oppenheimer’s (2006) to be an authoritative work in this rapidly evolving field. While a substantial portion of Oppenheimer’s Y-STR study also uses the Capelli data, he assigns new haplotype labels to the data – “16 distinct types of R1b” -- without ever providing the detail necessary to link his haplotypes to the underlying data. As is the case with Sykes’s work, this makes any real academic or juried review of his conclusions impossible and lessens the usefulness of his work to other researchers. In some cases like the aforementioned Pict test, researchers have been quick to partner with testing companies to make money from their private theories.⁴

The purpose of this paper is to take a look at the underlying Sykes R1b data and see if it can be linked to his founder haplotypes and the conclusions of his analysis. The goal of this paper is to attempt to provide additional insights to the work of these researchers to make it more useful to the individual genetic genealogist who look to their data as a link to the past.

Methods

This study focuses on Haplogroup R1b, which comprises the vast majority of the British Isles. R1b data was extracted from the targeted data set and two types on analysis were done to identify patterns – affinity analysis and network analysis. Network analysis was performed using the Fluxus Engineering’s phylogenetic network analysis software (Network 4.201). For the affinity analysis of haplotypes, an Excel spreadsheet was developed to look at patterns and anomalies of the data. An algorithm was developed that took as input the 10 marker values of

the haplotype signature and then compared this against the overall distribution of that haplotype in the full database. A sequential summary of the method is as follows:

1. Code Sykes’ Oxford Genetic Atlas Project (OGAP) data by haplogroup.
2. Separate the data into haplogroups and limit the analysis to R1b only.
3. Group similar R1b haplotypes.
4. Look for geographic patterns or anomalies in the distribution of major individual haplotypes.
5. Review the data “in the large” to see if individual patterns can be explained in a larger context and against Sykes’ published conclusions.

Results

Step 1 - Coding of the OGAP Data

For data collection, Oxford Genetic Atlas Project (OGAP) data was downloaded from Bran Sykes web site and converted from PDF to Microsoft Excel format. 2,322 samples were then coded by haplogroup using Whit Athey’s improved Bayesian haplogroup calculator (Athey 2005, 2006). Hereafter, “OGAP” will refer to the R1b subset of Sykes’s data.

Sykes included the following description of the data in the supplementary data file:

Y-chromosome DNA (yDNA) - Samples collected early in OGAP were amplified across the following seven markers: DYS 19, DYS 389i, DYS389ii, DYS 390, DYS 391, DYS 392, DYS 393 using conditions described by de Knijff et al (International Journal of Legal Medicine, 110: 134-140, 1997). Later samples were typed for these and three additional markers: DYS 388, DYS 425 and DYS 426, using the two-stage multiplex conditions described by Thomas et al. (Human Genetics, 105: 577-581, 1999). Alleles are reported as the number of repeat units. For reasons of continuity within OGAP, DYS 389i is reported as three repeats lower than the allele size produced by the ABI 3100. DYS 398ii-i reports the difference between 389i and 389ii, the reason being that the repeat size at 389ii is not independent of 389i whereas the difference between them is. Although Y-chromosomes were assigned to clades, largely by RFLP [Author - restriction fragment length polymorphism] analysis, these assignments are not reported here as they do not necessarily correspond to the SNP-based system recommended by the Chromosome Consortium (Genome Research 12:339-348, 2002).

Geographical distribution - Genetic data are assigned to geographical regions based on the

³ <http://www.ethnoancestry.com/Pictish.html>

⁴ <http://www.ethnoancestry.com/oppenheimer.html>

birthplace/residence of the paternal grandfather. This was done to minimize the effect of very recent migration. The regional boundaries are shown on a map which precedes the Prologue in Blood of the Isles. These data are copyrighted and must not be reproduced without permission. Other formats and additional details may be available for academic collaborations.

Several things are worth noting about the data. First, the Sykes data only uses 10 markers (DYS19, *DYS389i*, *DYS389ii*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS388*, *DYS425* and *DYS426*). In addition, only approximately 64% of the data are complete with 36% of the data missing four markers, *DYS439*, *DYS388*, *DYS425* and *DYS426*. While the missing markers appear to be a serious shortcoming, 94% of all the *DYS425* and *DYS439* markers and 73% and 74% of the *DYS426* and *DYS388* markers in Sykes full data set, respectively, have a value of 12.⁵ This means that these markers do not have sufficient spread and variability and are, in general, of limited use in discriminating between haplotype patterns within this set of data.

Another interesting fact is the haplogroup distribution of the data. Due to lack of haplogroup designations in the original data, Athey’s haplogroup calculator was used as a proxy to classify each haplotype. With several haplotypes being removed because of missing data, **Table 1** shows a comparison of the results of the haplogroup calculator with Sykes published “Clans.”

In this table, percentages shown in the middle column reflect the output of Athey’s calculator while the percentages in the far right column correspond to the breakdown published by Sykes in Appendix C of his book.

It is clear from this comparison that Athey’s calculator appears to classify the data in similar proportions as Sykes Clans and thus one may infer the underlying meaning of Sykes’ clan nomenclature.

In addition, it is important to note that Syke’s Clan categorizations were not based primarily on single nucleotide polymorphism (SNP) testing. As stated in the italicized quote above, “*Y-chromosomes were assigned to clades, largely by RFLP analysis, these assignments are not reported here as they do not necessarily correspond to the SNP-based system recommended by the Chromosome Consortium.*”

Finally, one should understand the regional borders that Sykes uses in his study. Since the purpose of this

analysis is to draw geographic inferences, we are limited in our insight by the definition of the geographic areas from which the data is collected. Figure 1 shows the regional borders that are coded in the OGAP data.

Table 1. Calculated Haplogroups vs. Sykes Clans

Distribution of Paternal Clans in England				
Haplo-group	Samples		Sykes Clan	
E3a	1	0.1%		
E3b	30	2.2%	Eshu	2.1%
G	17	1.3%		
I1a	188	14.0%	Wodan	22.2%
I1b1	32	2.4%		
I1b2	58	4.3%		
J	46	3.4%	Re	2.7%
Q	11	0.8%		
R1a	62	4.6%	Sigurd	5.2%
R1b	899	66.9%	Oisín	64.0%

Distribution of Paternal Clans in Scotland				
Haplo-group	Samples		Sykes Clan	
E3a	1	0.1%		
E3b	12	1.5%	Eshu	1.5%
G	2	0.3%	Other	0.3%
I1a	69	8.9%	Wodan	15.4%
I1b1	7	0.9%		
I1b2	32	4.1%		
J	17	2.2%	Re	1.2%
Q	5	0.6%		
R1a	67	8.6%	Sigurd	8.8%
R1b	563	72.6%	Oisín	72.9%

Distribution of Paternal Clans in Wales				
Haplo-group	Samples		Sykes Clan	
E3a	-	0.0%		
E3b	4	2.2%	Eshu	3.1%
G	7	3.9%		
I1a	11	6.2%	Wodan	11.0%
I1b1	1	0.6%		
I1b2	2	1.1%		
J	3	1.7%	Re	0.7%
Q	-	0.0%		
R1a	4	2.2%	Sigurd	1.4%
R1b	146	82.0%	Oisín	83.2%

Step 2 – Extracting R1b (Oisín) Data

The 1625 haplotypes identified as R1b in the previous step were extracted from the data set. These included all of the R1b haplotypes shown in Sykes Table 1, plus those for Ireland. It is also interesting to note that when only R1b is considered, the frequency of markers *DYS426*, *DYS388*, *DYS439* and *DYS425* with a value of 12 rises to 97% in this subset of the database. This further

⁵ For those markers that have been reported in the full data set. i.e., Alleles of 12 include: *DYS425* (1412/1496), *DYS439* (1411/1496), *DYS426* (1088/1496), *DYS388* (1108/1496).

reinforces the notion that these markers (or the lack of them) are not critical to the analysis of this set of data.

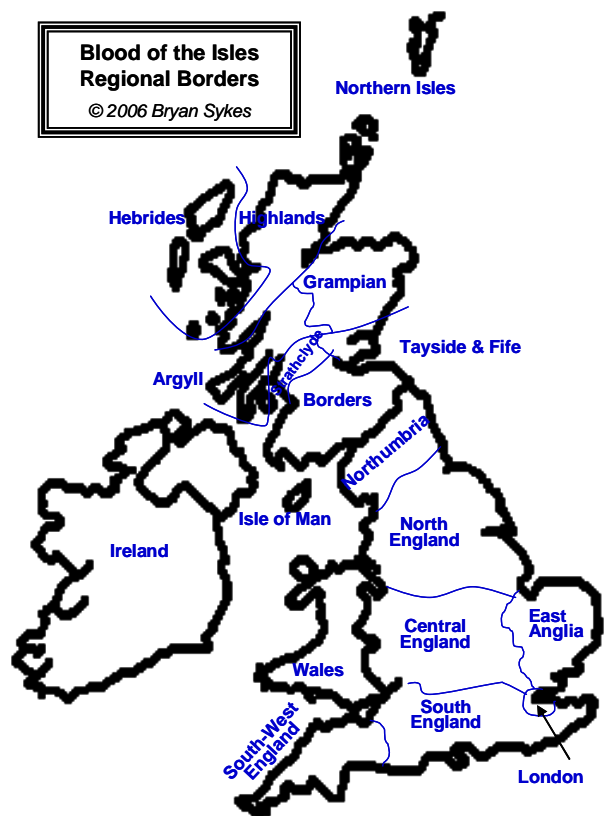


Figure 1. Regional Borders Used in the OGAP Analysis to Classify Individuals

The first thing that was done to better understand the data was to identify modal haplotypes for each region as a descriptive view of the R1b data set. However, examination of the modal haplotypes for the individual regions was not informative because all regions and the full data set matched the standard Atlantic Modal Haplotype.

A view of the R1b data in the form of a connected graph – as shown in **Figure 2** – shows a high degree of “cubism.” By this I mean a high degree of nodal interconnectivity among the data points that results in opposite vertices “washing out” differences in the data.

Clearly, data analysis based upon unique combinations of markers (i.e. haplotypes) instead of individual markers would be necessary.

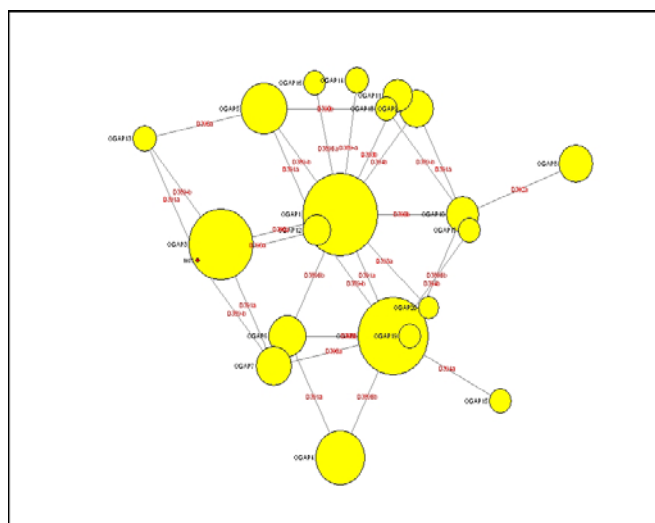


Figure 2. Network Analysis of the Top Twenty R1b Haplotypes

Step 3 – Analysis of Haplotypes

Since descriptive statistics tended to “average out” differences in the data, other methods were needed to identify patterns and analyze the data. To do this the most common haplotypes were identified and two methods of analysis were performed. Appendix A shows the haplotypes in the OGAP data. The OGAP haplotypes roughly follow the frequency distribution of YSearch and in McEwan’s (2007) groups if one takes into account that the OGAP data is light on Irish samples in comparison with these other sources.⁶

The OGAP designations in this table were assigned sequentially in decreasing frequency of occurrence. The OGAP numbers from this table will be referenced throughout the remainder of this report.

Two types of analysis were conducted to identify patterns – affinity analysis and network analysis.

For affinity analysis of the haplotypes, an Excel spreadsheet was developed to look for patterns and anomalies in the data. An algorithm was created that took as an input the 10 marker values for a haplotype or signature to be reviewed and then compared that haplotype to the R1b subset of the database. The algorithm calculated the genetic distance and reported back the number of perfect (i.e., zero distance) matches by OGAP region. To account for the differing level of sampling in each region (e.g., small for Ireland and large for Northern England), the resulting number of matches was normalized to frequencies in each region.

⁶ Only 22 of the 2322 samples (~1%) are from Ireland.

Table 2. Example of Identifying OGAP8 Affinities

REGION	R1b Data Set		Haplotype		Norm. Result
	R1b	%	Match	%	
Argyll	43	2.6%	4	11.8%	345.%
Borders	50	3.1%	1	2.9%	(4.%)
Central England	140	8.6%	1	2.9%	(66.%)
East Anglia	70	4.3%	0	0.0%	(100.%)
Grampian	106	6.5%	5	14.7%	125.%
Hebrides	52	3.2%	4	11.8%	268.%
Highland	60	3.7%	1	2.9%	(20.%)
Ireland	17	1.0%	2	5.9%	462.%
Isle of Man	25	1.5%	0	0.0%	(100.%)
London	92	5.7%	0	0.0%	(100.%)
North England	228	14.0%	1	2.9%	(79.%)
Northern Isles	119	7.3%	1	2.9%	(60.%)
Northumbria	58	3.6%	4	11.8%	230.%
South England	103	6.3%	0	0.0%	(100.%)
South-West England	183	11.3%	0	0.0%	(100.%)
Strathclyde	88	5.4%	6	17.6%	226.%
Tayside & Fife	45	2.8%	3	8.8%	219.%
Wales	146	9.0%	1	2.9%	(67.%)
	1625		34	100.0%	

For example, haplotype OGAP8 which is generally considered the quintessential Irish haplotype has 34 perfect matches in the database. Because Ireland was not extensively sampled, the OGAP database has twice as many hits for Grampian as for Ireland. However, when we adjust the Grampian and Ireland percentages for the nearly six times greater sampling of Grampian, we see that the greatest geographic affinity for the OGAP8 haplotype is Ireland. This calculation is shown in **Table 2** above.⁷ The OGAP8 results with decreasing presence in Ireland, Argyll, and the Hebrides also corresponds with what is generally understood to be the geographic distribution of this haplotype.

This analysis was repeated for the top 20 haplotypes in the OGAP data. These haplotypes, which cover 60% of all OGAP samples, appear sufficient to identify major regional affinities. Analysis of additional haplotypes would be increasingly subject to sampling error.⁸

The results of the analysis of the top 20 haplotypes are shown in **Table 3**. It should be noted that in this table negative values were removed to reduce the clutter and significant geographic anomalies were color coded to aid in identifying tendencies. Finally, especially interesting results were boxed to help in latter discussion in this paper.

⁷ The results column of Table 2 has the same relative weighting as if samples observed were divided by sample size (e.g. Ireland = 2/17) but this is just an alternative formula.

⁸ OGAP haplotypes below OGAP30 have single digit sample sizes.

License was also taken in the reordering of rows and columns of the table in an attempt to group similar haplotypes and close regions. While such analysis is called “affinity analysis” and can be conducted mathematically, this analysis was done manually to better allow for subjective considerations of the data.

The second type of analysis that was performed was network analysis. This analysis which is common in the genetic sciences was conducted using the Fluxus Networking program version 4.2.0.0. **Figure 3** shows the results of the network analysis for the top twenty OGAP haplotypes.

It should be noted in **Figure 3** that nodes have been relocated and line length changed for increased readability. Nodes have also been colored to reflect the regional affinities identified in **Table 3**.

Conclusions

Analysis of the Oxford Genetic Atlas Project data has yielded interesting results. The combination of the geographic affinity results shown in **Table 3** and the network analysis results shown in **Figure 3** are synthesized in **Figure 4**. In this graphic, key haplotypes with strong regional affinities were placed in their rough geographic perspective. No attempt was made to force every haplotype somewhere on the map as it is obvious that at this level of analysis, some haplotypes are pervasive and ubiquitous and not easily generalized to a single geographic region.

Once located, haplotypes that differed by a single mutation were connected with lines. **Figure 4** reflects

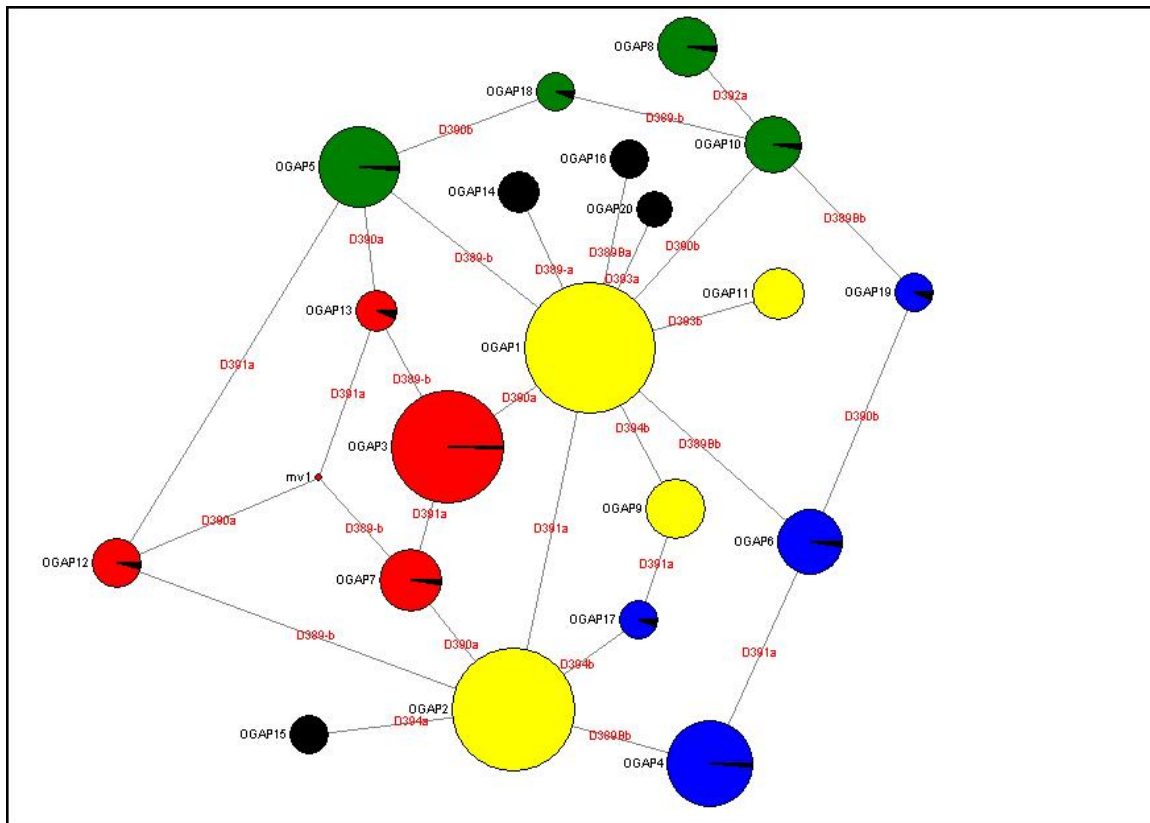


Figure 3. Network Analysis of OGAP Data

the general interconnectivity resulting from the network analysis of **Figure 3**. The lines in **Figure 4** should be thought of as one possible path of migrations – not necessarily the only path. The interconnections shown in **Figure 4** are not based on any individual mutation rates. The interconnections shown in **Figure 4** are based upon the occurrence of mutations, the principle of parsimony, and the general south-to-north flow of R1b discussed by Sykes and Oppenheimer. Parsimony, in this case, reflects the generally acknowledged flow from higher concentrations of haplotypes to lower, more diffused concentrations.

Some of the observations and conclusions of this analysis are as follows.

1. The methodology clearly identified and quantified what has been previously called the Irish subclade. Whether called the Irish Modal Haplotype or the “Ui Neill haplotype” as in the Trinity College paper⁹, the North-West Irish Haplotype by David

⁹ <http://vetinari.sitesled.com/gael.pdf>

Wilson and fellow researchers¹⁰, or R1bSTR19Irish as defined by McEwan (2007),¹¹ this corresponds to OGAP8 in the Oxford data. In addition, this analysis also identifies the strong migration of OGAP8 to the Argyll area. This migration may be explained by the Dal Riada migration¹² or as Sykes notes “*Ar-gael was coined by the kings of Dal Riada for their three colonies – Mull, Islay, and Kintyre peninsula.*” On page 217 he continues, “*There has*

¹⁰ <http://www.m222.net/R1b1c7> David Wilson also writes: “In 1,601 haplotypes tallied in Capelli’s data table, 60 show 25/14 at 390/392. Of these, 47 are of the 25/11/14 variety, and the remaining 13 show 10 at DYS391. These instances are geographically clustered in Ireland and along the west coast of Britain from Wales to the Orkneys. A few are spread out in Southern Scotland. With the exception of two instances in Norfolk, there are no instances of the pattern in England. Capelli also reports a small cluster in Norway (four instances) and a singleton in Germany/Denmark even though these locales are outside his major study area.”

¹¹ <http://www.geocities.com/mcewanjc/p3modal.htm>

¹² Mark McDonald writes, “*The Dal Riada (Dalriads) leadership who came from Ireland in circa 500AD into what is now Argyll spoke a language akin to what is now called Erse (Irish Gaelic to the Scots) and introduced that language into Scotland - the root of modern Scots Gaelic. They were called ‘Scoti’ by the Romans it is said - a word for ‘raider’ used in those days, and it is the root of the name Scotland which later developed as the tribes cohesed into a nation.*”

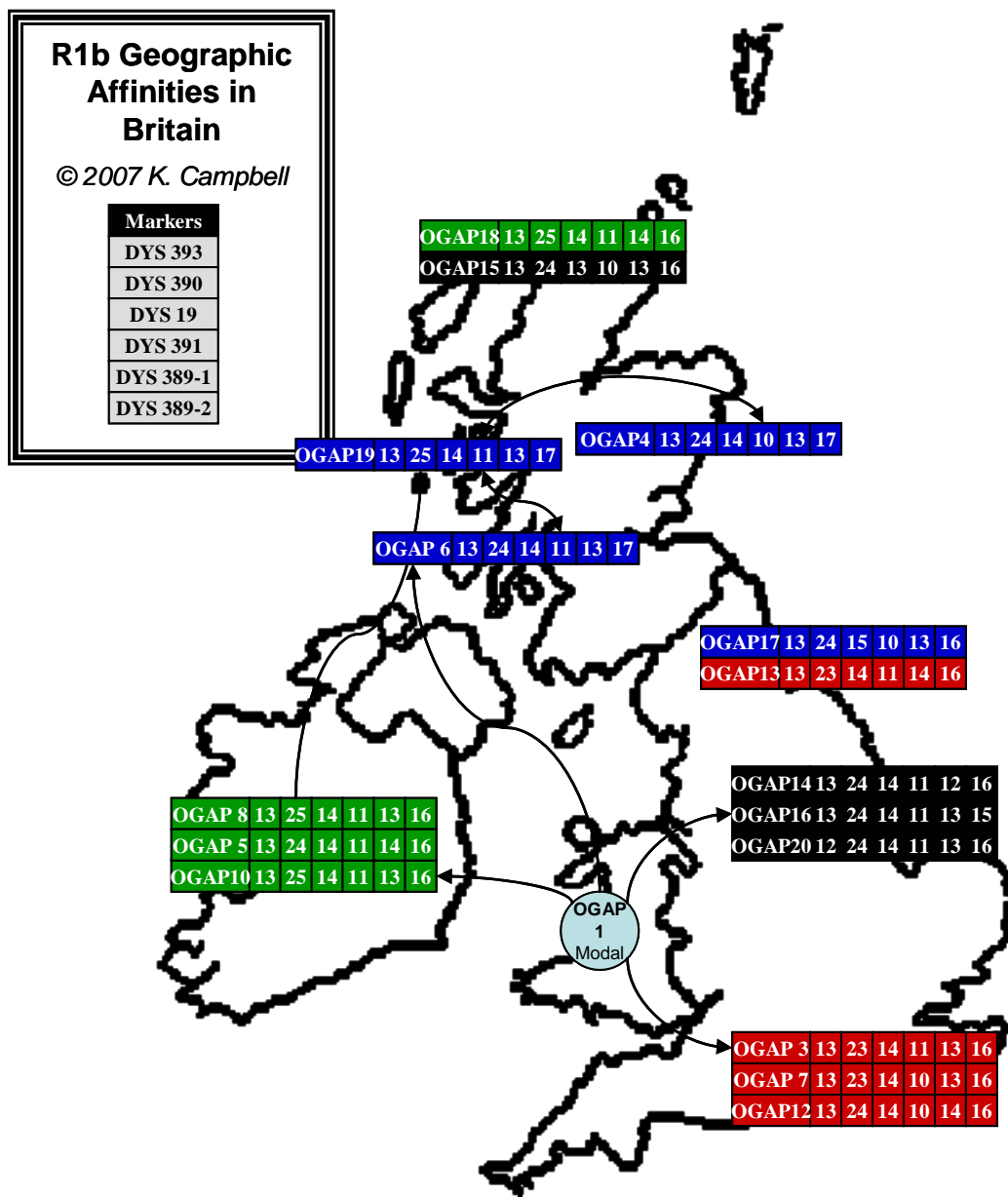


Figure 4. Geographic Patterns of R1b in the British Isles

certainly been a substantial settlement at some time from Ireland in the recent past and the Irish infiltration into the west of Scotland is almost certainly the signal of the relocation of the Dal Riada from Ulster to Argyll in the first millennium.”

2. Similarly OGAP10 which shares the DYS390=25 allele is clearly identified as an Irish haplotype. This haplotype also shows up strongly in the Hebrides which supports Sykes and Oppenheimer’s general conclusion of a Mesolithic northern migration along the coast of Ireland.
3. Interestingly, OGAP5 is a very prevalent haplotype that also shows up predominately in Ireland. What is interesting is that this haplotype has DYS390=24 instead of the generally accepted value of 25. This haplotype, which has only one difference from the AMH modal (DYS389-1 = 14) may be a transitional haplotype.
4. OGAP19 is interesting in that it shows an extreme correlation with both Ireland and the Scottish Highlands. Because of the low sampling of Irish participants, it is best to be cautious about drawing strong conclusions concerning Irish population genetics but clearly OGAP8, OGAP10, and OGAP19 show the strong ties that are shared by Ireland and the Highland area of Scotland. It is the migration of these haplotypes westward during the first millennium that make the Scottish to Irish migration in the 1700s so difficult to isolate genetically. Specifically Sykes writes, “*There is a basic similarity between Irish and Pictish Y Chromosomes which, incidentally, makes it almost impossible to detect any genetic effect of the Ulster Plantations.*”¹³
5. OGAP4 is particularly intriguing. It is ubiquitous across all areas of Scotland and exceptionally strong in Grampian, Tayside, and Strathclyde. If we discount the Irish influx in Argyll and the Hebrides, it is also among the strongest haplotypes present in these regions.¹⁴ It would not be too much of a stretch to label OGAP4 the quintessential Scottish haplotype and the single closest identifier to whatever is considered the indigenous Scottish population. Sykes and Oppenheimer both write that the Picts were as close as anything to the

¹³ Sykes, Blood of the Isles, page 210.

¹⁴ When writing about Argyll, Sykes writes, “*However, the genetic signal, as far as I can judge, points to a substantial, and by the look of it, hostile replacement of Pictish males by Dalriadian Celts, most of whom relied on Pictish rather than Irish women to propagate their genes.*” Sykes, Blood of the Isles, page 210.

indigenous population of Scotland.¹⁵ Sykes states on page 217,

6. OGAP6 is prominent in Argyll and the Hebrides. It is also probable that OGAP6 is what Sykes was referring to when on page 209 of his book he writes, “*The Argyll Y Chromosomes were much more like their Hebridean counterparts than those in the Highlands ... Argyll Y chromosomes are in between the Irish and the Pictish values... Argyll has a low gene-sharing score with Grampian, even after the Norse component has been subtracted.*” The first part of this statement can be observed by looking at the affinities for OGAP6. The second part of the statement further reinforces the notion that OGAP4 represents Pictishness since the OGAP6 (Argyll) is a connection of OGAP4 to the Irish constellation in Figure 3.
7. OGAP9 and OGAP11 both show an affinity for both the Northern Isles and the Borders regions. This affinity is distinctive, but the author is unqualified to venture a theory that might explain this geographic discontinuity.
8. OGAP13 and OGAP17 both show a clear affinity for Northern England.
9. OGAP14, OGAP16, and OGAP20 all show a common regional affinity. Though their presence in Tayside is very slightly higher than in Central England, the author believes that they are probably better categorized as English rather than Scottish. It is theorized that this signature probably migrated northward along the east coast of England, but the author is at a loss as to why there isn’t a greater presence of these haplotypes in the Borders, Northumbria, and Northern England. Perhaps more historically astute researchers might provide a theory.
10. OGAP7 seems to be most prevalent in Southern England. OGAP12 has a stronger presence on the Borders area, but similar to the comment above, it is

¹⁵ Other researchers suggest that this haplotype might be Dal Riadic Celt (see:

<http://searches2.rootsweb.com/cgi-bin/igetch2?/u1/textindices/G/GENEALOGY-DNA+2004+14773453734+F>).

However, the ubiquitous presence of OGAP4 across Scotland including its strong presence in the Pict areas of Grampian and Tayside tend to infer a more ancient introduction than the eastern movement of Celts to Argyll a thousand years ago. Sykes writes, “*There is no surviving mythology around the Picts.... Grampian and Tayside – is Pictland ... The reason I cannot be more certain is itself very relevant to the myth of the Picts. It is precisely because they are genetically close to the Gaelic Irish that these estimates are so difficult. If they had been a relic people, a genetic isolate, then it would have been easy to distinguish them from Irish Gaels. But on the contrary, it is extremely difficult, from which we can confidently conclude that the Picts and the Celts have the same underlying genetic origins.*” [Author – i.e., R1b)

speculated that this may have been a northward migration but again the author can't explain the apparent geographical jump.

11. The core haplotypes for the full British Isles are OGAP1, OGAP2, and OGAP3. By their nature these haplotypes are considered oldest and progenitors of the British R1b lines. This means that these haplotypes by their very nature should be, and are, much diffused. OGAP1 has a slight predisposition to Scotland and has its greatest concentration in Wales.¹⁶ OGAP2 has a slight affinity for Ireland, and OGAP3 has a slight affinity for Southern England. It is unclear if much more can be said about these haplotypes but with them one might see the original immigrants to the Isles and their possible paths along the western and eastern coasts.
12. Wales is interesting. Sykes on page 239 makes the statement that Wales has the lowest diversity of R1b variety than anywhere else in the Isles. However this analysis shows that Wales has no predominant haplotype and a relatively even distribution of other haplotypes used in this analysis.
13. Like Wales, the Isle of Man is at a similar crossroads. However, rather than the diversity of Wales, the Isle of Man shows a strong presence of three haplotypes. Looking at the geographic context of Figure 4, it can be seen that the Isle of Man was populated by OGAP10 eastward from Ireland and westward by OGAP13 from northern England and OGAP16 from central England.

Summary

Through the analysis of Sykes' OGAP data, this study has provided a means linking DNA results to haplotypes and conclusions in Sykes' book, "Blood of the Isles." The study has confirmed Sykes' interpretation of the data, and hopefully, provided a means for other researchers to further validate and extend his work. The study both confirmed some subclades identified by Sykes as well as identified some new subclades worthy of further research. Key subclades that the study posits and which are defined by Sykes include those of the Picts and the Dal Riada Celts.

Picts – It is asserted that OGAP4 best represents the Pictish ancestry of Scotland. While there is no fundamental genetic difference between the Picts and the Celts, i.e., both being R1b, they are both from the same

¹⁶ Again this analysis confirms the statement on page 239 of Sykes book, "The Atlantis Chromosome, the prevalent Y chromosome in the Clan, is very frequent in Wales, more so in Ireland, as a proportion of Oisins as a whole."

mixture of Iberian and European Mesolithic ancestry that forms the Pictish/Celtic substructure of the Isles.¹⁷

Dal Riada Celts – When considered in a narrow genetic sense, the Gaels of Ireland, as identified by the DNA signature of OGAP8, are as close as any group to being considered the root line and forbearers of Celts of today.¹⁸ When present in Scotland, it is suggested that OGAP8 represents the signature of the Dal Riada Celts. However, more generally, it should be clearly noted that the term Celtic has taken on a near mystical meaning and now more commonly refers to cultural, linguistic and regional patterns that have superseded any genetic definition. Sykes says it best when he states. "Overall, the genetic structure of the Isles is stubbornly Celtic, if we by that mean descent from people who were here before the Romans and who spoke the Celtic language. We are an ancient people ...genetically rooted in the Celtic past. The Irish, the Welsh, and the Scots know this, but the English sometimes think otherwise."

Several interesting clusters were identified that show geographic affinities but discontinuities. Scottish clusters OGAP9 and OGAP11 have a strong presence in the Borders as well as the Northern Isles. English clusters OGAP14, OGAP16, and OGAP20 show a predisposition to both Central England as well as Tayside. These and other geographic affinities warrant further research.

Electronic Database Information

Capelli, C. et al. 2003 Data Set:

<http://freepages.genealogy.rootsweb.com/~gallgaedhil/Capelli.htm>

Oxford Genetic Atlas Project (OGAP) Data Set:

<http://www.bloodoftheisles.net/results.html>

John McEwan's R1b Haplotypes:

<http://www.geocities.com/mcewanjc/p3modal.htm>

Whit Athey's Haplogroup Predictor:

<http://www.hprg.com/hapest5/>

¹⁷ Sykes, Blood of the Isles, page 282. It should be noted that Oppenheimer writes very little about the Picts in his book, Origins of the British. The reason may be that Oppenheimer's analysis based on the Capelli data relies only on 6 markers instead of 10. The lack of markers DYS439, DYS389-1, and DYS389-2 causes the Pict haplotypes (OGAP4) to be grouped with, and mask by, OGAP2 and OGAP12 in the Capelli data.

¹⁸ Sykes, Blood of the Isles, page 214 – "So far, we have four possible influences on the genetic structure of the people of Scotland, first the Picts, then the Gaels of Ireland, synonymous with the Celts, the Vikings, and in the south of Scotland particularly, the Anglo-Normans."

References

[Athey TW \(2006\) Haplogroup prediction from Y-STR values using a Bayesian allele-frequency approach. J Genetic Genealogy, 2:34-39.](#)

[Capelli C, Redhead N, Abernethy JK, Gratrix F, Wilson JF, Moen T, Hervig T, Richards M, Stumpf MP, Underhill PA, Bradshaw P, Shaha A, Thomas MG, Bradman N, Goldstein DB \(2003\) A Y chromosome census of the British Isles. Curr Biol 13:979-984.](#)

[McEwan J \(2007\) Phase 3 analysis: Ysearch 37 STR modal summary and analysis tables \(web site\).](#)

Oppenheimer S (2006) *The Origins of the British - A Genetic Detective Story*, Constable and Robinson, London (ISBN 1-84529-158-1). Published in the U.S. by Carroll and Graf, New York.

Sykes B (2006) *Blood of the Isles: Exploring The Genetic Roots of Our Tribal History*. Bantam Books. Published in the U.S. as *Saxons, Vikings, and Celts: The Genetic Roots of Britain and Ireland*, by W. W. Norton, New York.

Appendix A – OGAP Haplotypes

The 1625 data points that comprise the R1b data set include 291 separate haplotypes. However, 50% of the data can be accounted for with only 10 haplotypes; 60% by 20 haplotypes; and 68% by 30 haplotypes. In addition, haplotypes beyond the top 30 have only single-digit frequencies compared to the most frequent – the Atlantic Modal Haplotype (AMH), which occurs 262 times in the data.

Below are the top 50 haplotypes in order of descending frequency. These are number OGAP1 through OGAP55 for reference in this study and represent all haplotypes that occur more than 5 times.

The distribution of these haplotypes in Sykes' OGAP data and in YSearch (www.ysearch.org) as of December 2006) is shown on the right hand side of the chart.

Also, a mapping of John McEwan's (2007) R1b subclades is shown on the left hand side of the chart. The letters designating the McEwan group refer to the groups described in Appendix B that were assigned when McEwan individual haplotypes were grouped together to reflect the much smaller number of markers in the OGAP data.

McEwan (%)	McEwan Super Group	OGAP Haplotype ID	STR Marker and Value										Haplo-group	Sykes			Ysearch Count
			393	390	19	391	426	388	439	389-1	392	389-2		Count	%	Cum %	
38.9	(R1b Modal)	OGAP1	13	24	14	11	12	12	12	13	13	16	R1b	262	16.1%	16.1%	1818
14.7	E	OGAP2	13	24	14	10	12	12	12	13	13	16	R1b	138	8.5%	24.6%	767
11.6	C	OGAP3	13	23	14	11	12	12	12	13	13	16	R1b	117	7.2%	31.8%	746
5.2	(STR47Scots)	OGAP4	13	24	14	10	12	12	12	13	13	17	R1b	70	4.3%	36.1%	427
1.5	I	OGAP5	13	24	14	11	12	12	12	14	13	16	R1b	59	3.6%	39.8%	391
2.8	H	OGAP6	13	24	14	11	12	12	12	13	13	17	R1b	38	2.3%	42.1%	361
		OGAP7	13	23	14	10	12	12	12	13	13	16	R1b	35	2.2%	44.2%	286
7.2	(Irish)	OGAP8	13	25	14	11	12	12	12	13	14	16	R1b	34	2.1%	46.3%	511
		OGAP9	13	24	15	11	12	12	12	13	13	16	R1b	33	2.0%	48.4%	147
3.2	K	OGAP10	13	25	14	11	12	12	12	13	13	16	R1b	31	1.9%	50.3%	366
0.04	M (diff 1)	OGAP11	14	24	14	11	12	12	12	13	13	16	R1b	24	1.5%	51.8%	97
0.07	D	OGAP12	13	24	14	10	12	12	12	14	13	16	R1b	23	1.4%	53.2%	157
		OGAP13	13	23	14	11	12	12	12	14	13	16	R1b	17	1.0%	54.2%	159
		OGAP14	13	24	14	11	12	12	12	13	13	16	R1b	16	1.0%	55.2%	136
		OGAP15	13	24	13	10	12	12	12	13	13	16	R1b	15	0.9%	56.1%	18
		OGAP16	13	24	14	11	12	12	12	13	13	15	R1b	15	0.9%	57.0%	65
		OGAP17	13	24	15	10	12	12	12	13	13	16	R1b	15	0.9%	58.0%	93
		OGAP18	13	25	14	11	12	12	12	14	13	16	R1b	15	0.9%	58.9%	70
		OGAP19	13	25	14	11	12	12	12	13	13	17	R1b	14	0.9%	59.8%	52
		OGAP20	12	24	14	11	12	12	12	13	13	16	R1b	13	0.8%	60.6%	125
		OGAP21	13	22	14	11	12	12	12	13	13	16	R1b	13	0.8%	61.4%	7
		OGAP22	13	23	14	12	12	12	12	13	13	16	R1b	13	0.8%	62.2%	72
		OGAP23	13	23	15	11	12	12	12	13	13	16	R1b	13	0.8%	63.0%	78
		OGAP24	13	24	14	11	12	12	12	13	14	16	R1b	13	0.8%	63.8%	139
		OGAP25	13	23	14	11	12	12	12	13	13	17	R1b	12	0.7%	64.5%	87
		OGAP26	13	23	14	11	12	12	12	13	14	16	R1b	12	0.7%	65.2%	
		OGAP27	13	25	14	10	12	12	12	13	13	16	R1b	12	0.7%	66.0%	
		OGAP28	13	24	14	10	12	12	12	13	13	15	R1b	11	0.7%	66.6%	
2.5	B (diff of 1)	OGAP29	13	23	14	11	12	12	12	13	13	15	R1b	10	0.6%	67.3%	
		OGAP30	13	24	14	10	12	12	12	14	13	17	R1b	9	0.6%	67.8%	
		OGAP31	13	25	14	10	12	12	12	13	14	16	R1b	9	0.6%	68.4%	
		OGAP32	14	23	14	11	12	12	12	13	13	16	R1b	9	0.6%	68.9%	
		OGAP33	13	23	13	11	12	12	12	13	13	16	R1b	8	0.5%	69.4%	
		OGAP34	13	24	14	10	12	12	12	12	13	16	R1b	8	0.5%	69.9%	
		OGAP35	13	24	15	10	12	12	12	13	13	17	R1b	8	0.5%	70.4%	
		OGAP36	13	24	15	11	12	12	12	14	13	16	R1b	8	0.5%	70.9%	
		OGAP37	13	23	14	10	12	12	12	13	13	15	R1b	7	0.4%	71.3%	
		OGAP38	13	23	14	10	12	12	12	13	13	17	R1b	7	0.4%	71.8%	
		OGAP39	13	23	14	11	12	12	12	12	13	16	R1b	7	0.4%	72.2%	
		OGAP40	13	24	14	10	12	12	12	13	12	16	R1b	7	0.4%	72.6%	
		OGAP41	13	24	15	11	12	12	12	13	13	17	R1b	7	0.4%	73.0%	
		OGAP42	14	24	14	10	12	12	12	14	13	16	R1b	7	0.4%	73.5%	
		OGAP43	13	24	14	11	12	12	12	13	13	18	R1b	6	0.4%	73.8%	
		OGAP44	13	25	14	10	12	12	12	14	13	16	R1b	6	0.4%	74.2%	
		OGAP45	13	25	14	11	12	12	12	14	14	16	R1b	6	0.4%	74.6%	
		OGAP46	12	25	14	11	12	12	12	13	13	16	R1b	5	0.3%	74.9%	
		OGAP47	13	23	14	11	12	12	12	14	13	14	R1b	5	0.3%	75.2%	
		OGAP48	13	24	14	10	12	12	12	13	13	18	R1b	5	0.3%	75.5%	
		OGAP49	13	24	14	10	12	12	12	13	14	16	R1b	5	0.3%	75.8%	
0.8	G	OGAP50	13	24	14	11	12	14	11	13	13	17	R1b	5	0.3%	76.1%	

Appendix B – McEwan’s R1b Haplotypes Reduced

By mining YSearch and collecting similar 37-marker haplotypes into clusters, McEwan (2007) has identified a large number of R1b “types” that comprise the world-wide scope of this data. It is interesting to consider how the Sykes OGAP data relates to McEwan’s haplotypes. However to compare the data, McEwan’s modal haplotypes had to be collapsed into smaller groups to reflect the smaller number of markers in the OGAP haplotypes (Refer to Appendix A).

The following is the reduction of the McEwan haplotypes to 10-marker haplotypes. In each case, letter designations have been included here for the purpose of mapping the full set of McEwan haplotypes (designated R1bSTR##) to the reduced set used in this analysis (Letter Groups). While this exercise necessitates the loss of considerable resolution in the McEwan haplotypes, the exercise is included here to provide traceability to the analysis included in Appendix A.

Cum Freq	Super Group	McEwan Group	STR Marker and Value										
			393	390	19 394	391	426	388	439	389-1	392	389-2	
		R1b Modal (AMH)	13	24	14	11	12	12	12	12	13	13	16
1.1%	A	R1bSTR23	12	23	14	11	12	12	12	11	13	13	15
2.5%	B	R1bSTR24	13	23	14	11	12	12	12	11	13	13	15
11.6%	C	R1bSTR3	13	23	14	11	12	12	12	12	13	13	16
		R1bSTR20	13	23	14	11	12	12	12	12	13	13	16
		R1bSTR22Frisian	13	23	14	11	12	12	12	12	13	13	16
		R1bSTR25a	13	23	14	11	12	12	12	12	13	13	16
		R1bSTR27	13	23	14	11	12	12	12	12	13	13	16
0.7%	D	R1bSTR29	13	24	14	10	12	12	12	14	13	16	
14.7%	E	R1bSTR32	13	24	14	10	12	12	12	12	13	13	16
		R1bSTR35	13	24	14	10	12	12	12	12	13	13	16
		R1bSTR42	13	24	14	10	12	12	12	12	13	13	16
		R1bSTR43	13	24	14	10	12	12	12	12	13	13	16
		R1bSTR44	13	24	14	10	12	12	12	12	13	13	16
		R1bSTR45	13	24	14	10	12	12	12	12	13	13	16
5.2%	Scots	R1bSTR47Scots	13	24	14	10	12	12	12	12	13	13	17
8.2%	F	R1bSTR2	13	24	14	11	12	12	12	11	13	13	16
		R1bSTR5	13	24	14	11	12	12	12	11	13	13	16
		R1bSTR14	13	24	14	11	12	12	12	11	13	13	16
		R1bSTR18	13	24	14	11	12	12	12	11	13	13	16
		R1bSTR28	13	24	14	11	12	12	12	11	13	13	16
0.8%	G	R1bSTR33	13	24	14	11	12	12	12	11	13	13	17
38.9%	R1b Modal	R1bSTR1	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR4	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR6	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR7	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR8	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR9	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR10	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR11	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR12	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR13	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR15	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR16	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR17	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR26	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR30	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR36	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR37	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR38	13	24	14	11	12	12	12	12	13	13	16
		R1bSTR39	13	24	14	11	12	12	12	12	13	13	16
R1bSTR40	13	24	14	11	12	12	12	12	13	13	16		
R1bSTR46	13	24	14	11	12	12	12	12	13	13	16		
R1bSTR49	13	24	14	11	12	12	12	12	13	13	16		
2.8%	H	R1bSTR21	13	24	14	11	12	12	12	12	13	13	17
		R1bSTR34	13	24	14	11	12	12	12	12	13	13	17
1.5%	I	R1bSTR48	13	24	14	11	12	12	12	14	13	16	
1.2%	J	R1bSTR41	13	24	14	11	12	12	12	13	13	16	
7.2%	IRISH	R1bSTR19Irish	13	25	14	11	12	12	12	13	14	16	
3.2%	K	R1bSTR25	13	25	14	11	12	12	12	13	13	16	
0.4%	M	R1bSTR31	14	24	14	11	12	12	12	13	13	16	
100.0%													

Distance from AMH	0	1	2	3
-------------------	---	---	---	---