The Colla Phylogenetic Tree

Patrick McMahon M.Sc. ©
Mount Alexander
Gorey
Co. Wexford Y25 N799
Ireland
(*patmcmhn@gmail.com*)
July 2016

1. Introduction

The work described here draws together the considerable amount of SNP and STR data that has accumulated over the last few years with respect to the Colla population. Most recently, testing for SNPs has been encouraged using the advanced Big-Y tests (by FTDNA), targeted 'Pack' tests and single SNP tests. This has enabled the construction of inheritance Trees such as the one by Alex Williamson and the FTDNA Haplotree.

Separately and over a much longer period, STR data has been accumulating. These two (mutational) systems operate independently of each other to produce two separate data sets, the STR one defining members of the Colla population and membership of the Clan Colla 425 Null Project. As only about 27% of STR testers partook of SNP testing, the aim of this work was to populate/expand the resultant SNP Tree with STR data to determine the most likely clades for the majority of Project members.

This is an attempt to draw together the various findings collected to date (25 July 2016) on descendants of Clan Colla. All ancient genealogical data has been ignored, other than when the Colla brothers were thought to have lived and arrived in Ulster. I have concentrated solely on the DNA profiles and tried to rationalise their inheritance patterns and relationships. Surnames while confirmatory in many instances were not the main influence in determining groupings.

1.1. Pre-History

Long after man had migrated from the Rift Valley, humans had evolved into many different races and acquired different mutations thus making them distinguishable from each other. DNA studies have permitted us to categorise all humans on Earth into genealogical groups sharing one common ancestor at one given point in prehistory. These groups are called haplogroups.

Haplogroup sub-divisions are the result of very specific mutations whose times of occurrence have been calculated. Thus R1b-S116 branched into R1b-L21 by acquiring the additional mutation, L21. It is thought that a population of L21 people arose north of the Alps and spread through France, the Low Countries, Britain and Ireland starting about 4,000 years ago. It is estimated they could have started to populate Ireland about 3,000 years ago. By the time the Romans came to Britain, these Celtic people would have subsumed earlier cultures and diverged from one another over the 30 or so generations. They probably developed into tribal groups as they did in Ireland which the Romans, for their administrative convenience, gave those names such as the Trinovantes, Cornovii etc. Genetically, they would have similarities and differences as exhibited between Irish tribes.

1.2. Early Observations

The earliest method of classifying tester results was by comparing the GDs of name groups to the Colla modal and to each other. This resulted in useful cluster analysis for the more populous groups. However the main drawback was the huge level of cross relationships shown (expected within a population having a common ancestor) when viewed by GD alone making it difficult sometimes to assign testers correctly. As time went by it became clear that not everyone sharing a surname shared the same DNA and vice versa.

The onset of SNP testing has brought more rigour to how different lineages evolved. This introduced disparate groups, by name and STR profiles, being in the same clade. At first a bit of a conundrum but on reflection can be explained by either the SNP occurring before the STR profiles diverged or in other cases, in one of the diverged profiles only.

2. Definitions

The Colla population. The Colla population is defined as R-1b testers who have a null value (deletion) for the DYS 425 marker coupled with a value of 9 (or 10 in 2.8% of cases) for DYS 511. There are further confirmatory values of 9 at DYS 505 (99%) and 12 at DYS 441 (98%) for those who tested to 111 markers. More recently, having the Z3000 SNP or any of its downstream derivatives, additionally defines members of the Colla population.

The Colla name. Applying the name Colla to this population arose during early studies of Oriel clans where there was good documentary evidence of their descent from the Colla brothers. The early genealogy describes the appearance in Ulster in about A.D. 330 of the three Colla brothers, known as Colla da Crioch, Colla Uais and Colla Meann. Their arrival was allegedly at the behest of the High King of Tara. Their history survived through oral tradition and eventually written histories. It has since become apparent that they were not the originators of the Colla population but that the progenitor must have lived several hundred years earlier in Britain.

The Big Tree. Sterling work has been performed by Alex Williamson and Mike Walsh in analysing the huge quantity of data to emerge from Big-Y testing. This they have assembled into The Big Tree , a phylogenetic tree for the R-P312 haplogroup. The Collas are to be found within the R-DF21 major subdivision (P312>L21>DF13>DF21), one of whose sub groups (Airgialla 1) being for Z3000 (Colla) and descendants.

Airgialla 1. This Big-Y based Airgialla[1] sub-sub set is shown re-worked with additional information in the 'SNP Tree' worksheet of the Colla Phylogenetic Workbook [2] augmented with FTDNA Haplotree Z3000 batch and single SNP results. As Airgialla is the collective name of a territory occupied by a confederation of tribes, the term Colla is used to describe the genetically identifiable population in this analysis. The Clines[3] and nested Clades[4] within the population are defined by the most terminal SNP, either by name or chromosome address. The 121 SNPs identified so far (25 July 2016) were regarded as 'anchors' for their clades. This was the starting point of the current study.

2.1. Data Sources

The SNP source used was the Airgialla 1 Tree (described above). To this was added STR data as collected in the Clan Colla 425 Null Project (the Project). Only the data for 67 and 111 marker tests were used, exceptions being for those who had an SNP result but less than 67 marker tests.

2.2. Methods

The method developed was essentially one of pattern matching, substitution and iteration. As the Colla population descended from the DF21 population, neither modal[5] value could be used to distinguish clades or sub-clades from each other. Consequently, off-modal values were sought for each tester's STR results which were referred to as 'Discriminant' values for that tester's profile. Starting with existing clusters (based on surnames), the most representative discriminants were selected to define the genotype of

[1] The Kingdom of Airgíalla (Anglicized as Oriel) was one of the three major kingdoms that formed what is now the province of Ulster.
[2] The workbook is an OpenOffice calc document and you may need to download the Open Office program (free) from Apache Open Office in order to see it correctly.
[3] Biocommunities which display a continuous gradient. Genetically, clines result from the change of allele frequencies within the gene pool of the group of taxa in question.
[4] A clade is a group of organisms that consists of a common ancestor and all its lineal descendants and represents a single branch on the 'tree of life'.
[5] Modal values were determined by the most numerous groupings and by definition were values present in more than 50% of the sample. Modal values are most useful for identifying with the population as a whole but not in identifying sub-groups within.

the clade or sub-clade being assessed. The aim was to find up to 4 discriminants to define each sub-clade.

The SNP Airgialla 1 tree was the clade structure to which was added STR clusters that contained the SNP testers. The representative STR genotypes were used to search the tester database to identify those to be added to the emerging clades. Close matches to the genotype without an SNP could be validly added provided their GDs were low to the rest of the clade.

2.3. Assumptions

The driving assumption behind this analysis was that both SNP and STR mutations arose independently of each other. Over time, patterns of mutations would have built up (in separate branches), these patterns being inherited (largely unchanged) on the Y chromosome down through the generations. There would be some loss of discriminant values due to reversion and some gains due to new mutations.

Today's tester values represent the end point of this process so a tester with 4 (or 3 or 2 in descending order) discriminants will be deemed to have inherited this pattern with high probability rather than acquire it through random mutation. It was further assumed that should the cluster include a tester(s) with a defined SNP the whole cluster could then be classified as belonging to that clade.

In order to assess how many sub-clades were in a given clade, use was made of the McGee Utility Tool using the hybrid mutation model setting. Clade members were ranked by GD (using the McGee levels of significance) and where sub-clusters could be identified, these were classified as sub-clades and new representative genotypes identified.

2.4. Limitations

The main limitation was that there was a dearth of discriminants for some clades which could often be compensated for by having a well developed cluster structure. Another problem was there were a few clean clusters which had no associated SNP defined clade. Likewise, a small number of singletons could not be placed presumably because they were members of as yet undiscovered clades.

As this is very much a work in progress, the output will be subject to modifications and upgrades over time. In order to facilitate this I will upgrade my Dropbox copy at irregular intervals while retaining the links given here.

3. Results

The end results of this analysis are shown in the 'Clade Tree' worksheet (of the Colla Phylogenetic Workbook) and details of how the Clade Tree was derived are given in Appendix 1.

In the 'Clade Tree' worksheet, the Tree is displayed horizontally coupled to the STR data for eligible Project members.

3.1. Structure

Row 5 shows the mutation rates (where known) for the 111 STR markers (M Heinila, from K Nordtvedt, 2012), the darker the shading the more stable the marker; this helps in evaluating the weight to put on discriminant values. Each (tester) row of STR data was compared to the Colla and DF21 modals (rows 7 & 8) to find the off-modal values which are colour highlighted to show differences in the number of repeats, green = 1, yellow = 2, red = 3 or >, between them and the DF modal (except for the 4 Colla key marker values).

The banner heading colours conform to the 'SNP Tree' worksheet and show the clade name (most terminal SNP), the cline or lead SNP for nested clades in square brackets and the evolutionary path for the clade. The remaining headings are shown in Table 1.

Column(s)	Function	Comment
A-I	Tree structure	Shows descent of clades from DF21.
I	Number of SNPs detected	Totals per clade
J	Clade names	Mainly derived from the Big Tree but also showing alternative listings by FTDNA.
K	Sub-clades/ Genotypes	Branches from a clade showing distinct genotypes.
L	Lead reference (Ref) for clade	Most often SNP from Big Y but sometimes another member of the cluster.
L	Sub-clade reference Ref with GD to lead reference (#)	Visible as sub-clusters in McGee and having a different genotype to the lead reference.
L	Values (colour coded) are GDs to preceding Ref	Copied from McGee output.
M	ID Magenta = Big Y ID Blue = Pack test ID Green = Single test	Possible members are shown greyed (2 or < discriminants, high GD) or amber (no discriminants, compatible GD).
N-DK	DYS markers	Number of repeats under each marker.
DL	Count for each banner heading	Totals in DL598 & N598. Other totals in row 598.
DM-DQ	Transferred information from analysis worksheet.	Discriminant cells whose values used to match with genotypes.
DR-DS	# discriminants within 67 markers & # of hits	Preference given to discriminants within 67 markers.
DT	Max = highest possible score	Usually for 67 markers only.

Table 1. Column headings for worksheet 2.

3.2. Genotypes

Eighty two genotypes were identified out of 439 testers (indicated by a 1 in col DU of the 'Clade Tree' worksheet) of which 63 had a designated SNP. Up to 4 off-modal values (discriminants) were selected to represent each genotype. Some genotypes had more than four values which could be used to represent it while others had fewer than four. Each genotype was compared to every tester in the sample in order to detect likely matches. Initially, this was based on the number of discriminant hits, visually adjusted should the discriminants be in a different order and then examined in McGee. Should the clade show nested clusters then this was indicative of sub-clades.

3.3. Colla Relationships

In building the 'Clade Tree' worksheet, effort was focussed on maximising the number of close matches within each category. By subdividing the clades into sub-clades either because of the presence of more than one SNP which was different or where several clusters were apparent (within the clade as observed in McGee), it was possible to get tight relationships for most testers. The tight relationship (high probability) was generally indicated by GDs of 6 or less (@67 markers) to the preceding reference. Low probabilities were either GD > 6 and/or designated as 'Possible'. This resulted in the following assignments within the Clade Tree (Figure 1).

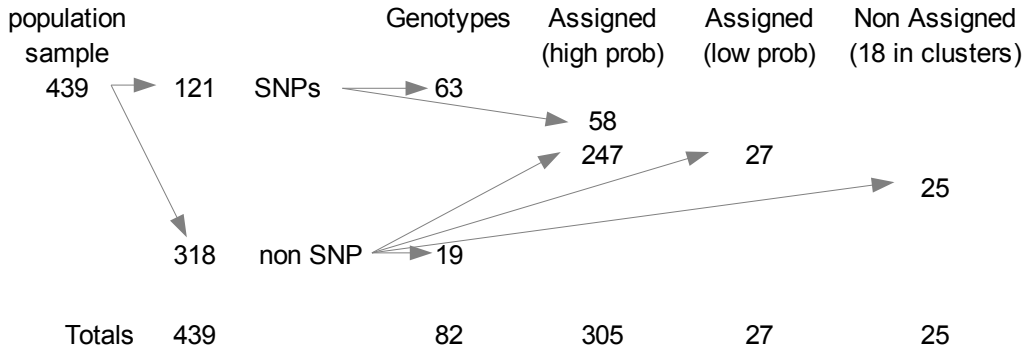

Figure 1. Assignment of the Colla Population Sample.

3.4. Colla Stem structure

The evidence (so far) points to the SNP Z3000 as the patriarchal ancestor of the Colla stem on the R-P312 Tree (which of course is part of the much bigger R1b tree). The clade structure can be seen in the 'SNP Tree' worksheet but can be difficult to track. A condensed version is shown in the 'Clade Structure' worksheet for the major components and where the numbers of testers fall.

The largest group descends from Z3004 at 287 testers. These further divide into 101 from S953, 93 from Z16274, 40 from BY3163, 14 from Z37586,12 from FGC41930 and 27 from Z3004 itself.

The next largest group descend from Z16270 at 106 testers. These further divide into 66 from BY3165, 27 from Z18003 and 10 from Z16270 itself.

The least represented group are the 7 from Z29586.

The unassigned comprise 4 named clusters of 18 testers and 7 singletons. It is a fair assumption that these are representatives of as yet undiscovered clades or sub-clades. SNP testing should resolve the issue.

4. Discussion

This analysis is based on ‘ownership’ of SNPs and key off-modal STR values (Discriminants) rather than differences in the numbers of DYS mutations between individual profiles and a reference modal. In performing this type of analysis, it was assumed that ownership of (named) SNPs produced a stable skeletal tree to which could be validly added profiles (without an SNP) which met the matching criteria. It could be argued that high STR mutation rates, which most discriminants have, would produce a confused picture. However, as the emphasis was on matching up to 4 discriminants collectively lessened the adverse effect of high mutation rates. The probability of multiple values being inherited as a group far exceeds them arising spontaneously. For example, DYS464 is recognised as unstable and yet the configuration of 15-15-15-17 is virtually a signature for the nested clades under Z16277. When combined with a second discriminant DYS557 = 17 (having a moderate mutation rate), and a third DYS522 = 12 (for 111 marker testers), produces unassailable evidence of relatedness.

Some clades are very clear cut, e.g. 6734552-A-G and Y9435 with as many as 6 discriminants. FGC41930, while only having 2 discriminants in 67 markers (and one extra @ 111 markers) is nevertheless unambiguous. Importantly, it demonstrates the justification for establishing sub-clade (2) which shares the 2 discriminants from (1) but because it has additional discriminants, it has a different genotype. This pattern is repeated often e.g. Z21270 (1) & (2), BY3162 (1& (2), Z18003 and its nested 19219990-A-T, BY3249 (1-3).

Perhaps the clearest examples of familial inheritance are the nested clades under A938. DYS 19 (a fairly stable marker) having a value of 15 was present in S953 and subsequent clades. This was joined by DYS534 = 17 from A938 onward and by DYS437=14, DYS458=18 from A939 onward. This is an expected result as the bulk of testers were McDonalds or derivatives and would be expected to show these types of familial relationships.

Like all biological systems, there are a few low probability assignments classified as ‘possible’. These lack SNP data and could fit into more than one clade; their current positioning was judged to be the most appropriate but could change in the future. Similarly, the 7 unassigned were so unrelated to the tester data set that they didn’t even qualify as ‘possible’. Finally, there are 18 grouped into 4 clusters which show no associations. All these uncertainties could be resolved by taking the Z3000 Pack tests.

This process of analysis has been developed over many months during which time it has been challenged every time new clades &/or data have been introduced. It has accommodated these changes well and would appear to be reasonably robust and a valid approach to phylogenetic tree analysis. Undoubtedly it could be improved upon by developing a ‘macro’ to handle the various pattern matching tasks.

5. Conclusions

In conclusion the following can be stated for the Colla population:

- The analysis has produced a comprehensive phylogenetic tree accommodating 88% of testers with high probability, 6% at low probability and 6% unassigned.
- The SNP structure has allowed STR profiles, previously regarded as unmatched, to be accommodated in the same clade.
- Clades can be divided into sub-clades on the basis of their SNP membership and/or the STR profiles.
- The unassigned and low probability matches could be resolved with SNP testing.

6. Appendix 1

The detailed workings are given in the 'Clade Workbench' worksheet (of the Colla Phylogenetic Workbook). Over the course of this study 82 genotypes were identified to represent the defined clades and sub-clades and 4 for the unassigned clusters. These are shown in rows 5-90 of the worksheet and the STR results in columns N-DK. Qualified testers are listed, initially in any order, in cells I101-DK101 onward.

In order to find testers which matched the genotype under test, conditional true/false equations were developed for 1, 2, 3 & 4 discriminants. These are shown in columns DL-DO for each genotype. The spreadsheet cells containing the selected discriminant values for each genotype are in columns DT-DW. Column DX shows the number of discriminants within 67 marker tests. These specific cells have to be used to allow the equations to work.

The data was processed in the following manner:

1. A genotype was copied (cells I#-DX#) and pasted into cell I100.
2. Select DL100-DO100 and 'Fill Down' against all testers.
3. Look for the maximum numbers of hits and record result in columns DP & DQ. Overwrite lower scores where appropriate.
4. When all genotypes have been tested, do a data - sort for column DP (ascending) and DQ (descending).
5. Copy the data blocks into the Clade Tree (discriminants and hits in columns DM-DT).
6. Visually inspect to recover missed hits (because an early discriminant was negative).
7. Using McGee, rank individuals in each clade/sub-clade (col L).

6.1. Worked Example

In the 'Clade Workbench', one sub-clade (out of 5), BY2869 (2), from a fairly typical but difficult clade, is examined. Its genotype is highlighted (row 57, light grey) and copied into I100. The results are shown in rows 243-247.

This is a strong result with 2 testers scoring the maximum of 4, 1 tester a maximum of 3 (67 markers only) and 2 being increased to 3 because the 3rd discriminant was negative but the 4th positive.

The results also show the familial connectivity of all the sub-clades.

6.2. McGee Results

This is further endorsed in McGee when the STR values for the whole clade are entered (Figure 2). The GDs from the lead Ref to the other 4 sub-clades are given as 6, 8, 8, 10 respectively. The GDs in turn to the sub-clades are shown horizontally (Table2 &col L).

Sub-Clades	GDs
263699(1)	6, 6
263699(2)	4, 4, 5, 6
263699(3)	1, 3
263699(4)	3, 4, 4
263699(5)	5

Table2. BY2869 sub-clade GDs

Genetic Distance																		
ID	263699Conly	11189Cavender	78625Conley	253386McCormick	169972Mcconnell	311107Mcconnell	84262Conley	328903Conlee	83647Dever	60356Smith	74097Smith	N18664Alexander	H2133Henson	342831Conley	338710Conley	N88161Connolly	B19040Connolly
263699Conly	-	6	6	6	8	8	7	8	8	9	9	8	7	8	8	10	11
11189Cavender	6	-	4	6	8	8	7	8	8	9	9	8	7	8	8	8	9
78625Conley	6	4	-	6	8	8	7	8	8	9	11	8	7	8	8	8	9
253386McCormick	6	6	6	-	4	4	5	6	6	7	9	8	7	8	8	8	9
169972Mcconnell	8	8	8	4	-	1	5	6	7	8	10	9	8	8	9	8	11
311107Mcconnell	8	8	8	4	1	-	6	6	6	7	9	8	7	8	8	9	11
84262Conley	7	7	7	5	5	6	-	5	6	7	9	8	7	7	8	7	10
328903Conlee	8	8	8	6	6	6	5	-	6	7	9	6	7	6	8	4	9
83647Dever	8	8	8	6	7	6	6	6	-	1	3	8	7	8	6	9	11
60356Smith	9	9	9	7	8	7	7	7	1	-	4	9	8	9	7	10	10
74097Smith	9	9	11	9	10	9	9	9	3	4	-	11	10	11	9	12	14
N18664Alexander	8	8	8	8	9	8	8	6	8	9	11	-	3	4	4	7	11
H2133Henson	7	7	7	7	8	7	7	7	7	8	10	3	-	3	3	8	10
342831Conley	8	8	8	8	8	8	7	6	8	9	11	4	3	-	4	8	11
338710Conley	8	8	8	8	9	8	8	8	6	7	9	4	3	4	-	9	11
N88161Connolly	10	8	8	8	8	9	7	4	9	10	12	7	8	8	9	-	5
B19040Connolly	11	9	9	9	11	11	10	9	11	10	14	11	10	11	11	5	-

Figure 2. McGee output for the BY2869 clade @67 markers, hybrid mutation model. Colour coding as for McGee.

6.3. Heterogeneity of Clades

Sizeable clades (>4 members) invariably sub divided into sub clades but a few were homogenous such as BY513, BY3164, A945, 6734552-A-G, Y9435, BY3166, 23753717-C-A. Heterogeneous clades were broadly of two types, those whose sub-clades were designated by a shared SNP value but different discriminants and those who shared some of the discriminant markers but were sub-clusters (in McGee).

6.4. Names

A general observation was that many of the clades contain very different surnames. Many of these are present because they have SNP anchors carrying the different name in that clade. Examples are Owens, Hines and Baty in 14095734-G-A; Johnson, Rogan, Highlands, Henretty in BY3165. Others have a mixture of familial continuity and an SNP anchor (Roderick, Clark, Adams in Z21270). Yet others are based purely on familial continuity (Calkins, Plunket in Z18003).

As most of these SNP anchors have large GDs to the Ref, a probable explanation is that the divergence in SNP terms arose long before the introduction of surnames. For those with smaller GD differences, the likely explanation is that the SNP lineage diverged in STR terms but still probably before the introduction of surnames (e.g. Clark (GD=7), Plunkett (GD= 6), cited above).

6.5. SNP Classification

A problem was encountered in applying SNP names to the clades. Choosing the lead SNP (as practiced by Alex Williamson) was followed as far as possible. The problem arose in trying to combine Big-Y, pack and single test results from FTDNA with those on the Big Tree. The 'SNPs @ 25-7-2016' worksheet lists the equivalents to aid translation.