The history of M222

A story in six parts

Iain Kennedy

query@kennedydna.com

July 2014

Part One. Discovery.

Ironically the early story of M222 was centered around a medical study into male infertility. A group of scientists in the United States carried out a study using DNA samples from an infertility clinic in St. Louis, Missouri along with a control set of fertile men. By examining certain genes in these mens' Y chromosomes a number of new mutations were found in genes related to fertility. One of these genes went by the cumbersome acronym USP9Y (originally called DFFRY). These mutations were then further examined to determine if they actually effected fertility. One, a G to A base transition, was confirmed in the discovery paper as having no medical impact (fertile men had the mutation too) and was named USP9Y 3636. This discovery was reported by Sun et al. in 1999[1] and co-author Page subsequently dubbed the mutation PAGE00084. A year later another geneticist Peter Underhill released a survey paper of Y chromosome mutations useful to population studies[2] and all (re-)christened with his prefix code M; hence the new name M222 for USP9Y 3636/PAGE00084, citing the Sun et al. paper of the previous year. Since then the shorter name M222 has stuck and the original names are all but forgotten.

Despite Underhill using M222 in a study of phylogeography without apparent reservation, the 'infertility' tag stuck for several years and as late as December 2005 Dr. David Faux of Ethnoancestry was recommending avoiding the marker for ancestry testing[3].

Part two. 'North west Irish' and Niall Noígíallach.

By the time of Faux's remarks above, the amateur (or 'citizen-scientist') field of genetic genealogy was in full swing, and in its early years focussed around a Rootsweb genealogy list 'genealogy-dna'. SNP testing was almost unknown partly due to the lack of suitable markers and testing was typically based around STRs (short tandem repeats). Early testers amongst the Irish and Scottish frequently found a common pattern with lots of close matches. One researcher David Wilson spotted a

---

[1] Sun et al. An azoospermic man with a *de novo* point mutation in the Y-chromosomal gene USP9Y. Nature Genetics volume 23 429-432 (1999.).

[2] Underhill P. et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Annals of Human Genetics Volume 65 Issue 1, 43-62 (2001).

[3] Faux D., Rootsweb post at http://archiver.rootsweb.ancestry.com/th/read/genealogy-dna/2005-12/1135329730, 23 Dec 2005.

common pattern[45] of STR markers with a strong bias to Ireland and western Scotland. His early comments albeit speculative may have been quite prescient, invoking carriers moving up the Irish Sea dividing left and right.
It was a full year later when a defining scientific paper from Trinity College Dublin was published, dubbing the same Wilson haplotype 'Irish Modal haplotype' or 'IMH'.[6] Needless to say, and not for the first time in this saga, there was no reference to the amateur discovery of a year ago which was already in the public domain. This was a truly seminal paper (Google Scholar reports 70 citations as of July 2014) which captured the public imagination by linking the IMH to the progeny of Niall Noígíallach, arguably the earliest truly historic figure in Irish history and founder of the Uí Néill kindred. The moniker IMH never really caught on at least in amateur circles with most people preferring 'North west Irish' or simply 'Irish', the latter being the term deployed for comparison by Wilson when the Moore paper was released:

> "*The TCD team's IMH is the same thing as the 25/11/14 "Irish" variety (McEwan's STR19) that we discuss on this list.*"[7]

It is worth pointing out that the academic paper was the result of a long field study and had gone through the standard peer review process. The amateurs could publish their results much faster although Patrick Guinness reported the completion of the draft of the study, without any details, two months before Wilson publicly identified the haplotype[8].

A detailed critique of the Moore paper is outside the scope of this article; two recent academic reviews can be found in Duffy (2013)[9]. A combination of high level SNPs including R1b-M269 (but not M222) and a small set of 17 STRs were typed from all the samples; it is a fortunate fact of M222 history that the distinctive STR pattern is visible amongst the most commonly tested markers (DYS390-392). The primary samples were only collected in Ireland but a comparison was made with British data from Capelli's recent study[10]. This led the authors to note a pattern of IMH in 'western and central Scottish locations'; noting historical links between the north of Ireland and Scotland they did not, in fact, make a statement about which direction the IMH might have flowed between the two.

---

[4] Wilson, D. '390=25 and 392=14 in Capelli and other data sets', Rootsweb post http://archiver.rootsweb.ancestry.com/th/read/GENEALOGY-DNA/2004-12/1101950215 1 Dec 2004
[5] Wilson, D. 'Defining a new R1b variety: 25/11/14' Rootsweb post http://archiver.rootsweb.ancestry.com/th/read/genealogy-dna/2004-12/1102007091 3 Dec 2004.

[6] Moore, L. et al, 'A Y chromosome signature of hegemony in Gaelic Ireland', American Journal of Human Genetics 78/2 334-338 (2006).
[7] Wilson D., Rootsweb post http://archiver.rootsweb.ancestry.com/th/read/genealogy-dna/2005-12/1134226716 10 Dec 2005
[8] Guiness, P., 'Irish clans' study / progress update' Rootsweb post http://archiver.rootsweb.ancestry.com/th/read/genealogy-dna/2004-10/1098012825 17 Oct 2004
[9] Jaski, Bart. Medieval Irish genealogies and genetics. Chapter 1 pp3-17 in: Princes, prelates and poets in medieval Ireland. Essays in honour of Katharine Simms, ed. Sean Duffy (Dublin, 2013);
Swift, Catherine. Interlaced scholarship: genealogies and genetics in twenty-first century Ireland. Chapter 2 pp18-31 in Duffy, op. cit.

[10] Capelli C. et al., 'A Y chromosome census of the British Isles', Curr Biol 13:979-84

In order to investigate whether the high concentration in north west Ireland might be associated with the Uí Néill they then sampled 59 men with Uí Néill surnames. Noting a strong correlation, they attributed '**a rise in frequency**', but not the origin, of the IMH to the success of the Uí Néill.

"Genealogical association together with the predominance
and pattern of variation of the IMH strongly suggest
a rise in frequency due to strong social selection
associated with the hegemony of the *Uı´ Neill* dynasty
and their descendents."

The attraction of this narrative, unsurprisingly stripped of its relative caution and caveats, was a boon to several parties not least of which were commercial testing companies who saw a lucrative market in selling a test kit which would show the customer was a descendant of Niall Noígíallach – even though the paper said nothing of the sort[11].

Part three. M222 meets the NWI/IMH haplotype

It was David Wilson again who now turned his attention to finding the SNP that defined the NWI/IMH haplotype and he first posted his idea that the former was M222 on the Rootsweb genealogy-dna list in February 2006[12], at a time that no commercial company actually sold the M222 test. Shortly afterwards, Ethnoancestry bowed to customer pressure and released the M222 test commercially.[13]. The author was a day one participant and was quickly confirmed derived for M222. As more testing was conducted it was quite evident that the STR-SNP link was correct and in fact M222 is still highly predictable from STR patterns[14].

Once again the scientists were not far behind, at least measured by article publication. Dennis Garvey had his paper linking M222 to IMH published later in 2006[15] and again, no mention was made of Wilson even though Garvey was a regular contributor to the mailing list where Wilson was posting. The original TCD team of Bradley, Moore and McEvoy seem not to have returned to the subject, although during Bradley's work on the first Irish genome sequencing in 2010 it was referred to in passing as being ancestral in the test subject[16] and they recapped the study in the 2012 book 'Celtic from the West'[17].

---

[11] Family Tree DNA, 'Matching Niall Noigiallach', http://www.familytreedna.com/landing/matching-niall.aspx

[12] Wilson, D., 'Is the NW Irish variety actually R1b1c7 (M222+) ?',
http://archiver.rootsweb.ancestry.com/th/read/genealogy-dna/2006-02/1140413475 19 Feb 2006
[13] Faux, D., 'Ui Neill Irish lineage SNP located', Rootsweb post
http://archiver.rootsweb.ancestry.com/th/read/genealogy-dna/2006-03/1141526628 4 Mar 2006

[14] Casey, R., 'Analysis of M222 SNP',
http://www.rcasey.net/DNA/R_L21/Analysis/R_L21_Analysis_M222.html, April 2013
[15] Garvey, D., 'Sub-Populations Within the Major European and African Derived Haplogroups R1b3 and E3a Are Differentiated by Previously Phylogenetically Undefined Y-SNPs', Human Mutation: Mutation in Brief #940 (2007).
[16] Tong. P et al, 'Sequencing and analysis of an Irish human genome', Genome Biology 2010 11:R91

Part four. 2007 to 2012, the long wait?.

An assessment of progress between the establishment of M222 as the IMH SNP in 2006 and the eventual subdivision of M222 via next generation sequencing depends on the reader's attitude to STR-based analysis. The challenge of making progress was hampered by the geographical closeness of Scotland and Ireland and the many known population movements between the two, together with an incomplete understanding and indeed ambiguity concerning the true origins of surnames. Another problem was the quality of data available. Amateur data often consisted of a rich selection of STR markers but with poor or non-existint geocoding information and highly skewed sampling, whereas academic papers had rigorous sampling but lagged badly behind in marker selection. Eliminating selection bias from commercial database samples was a big challenge and arguably one of the best attempts was the Conroy paper which in surveying all regions in Britain and Ireland concluded that south-west England demonstrated the highest diversity and was therefore the probable place of origin of M222. An interesting piece on the presence of M222 in Galloway has been produced by Milliken and is being updated based on new genetic data. An interesting academic study into Viking DNA in the Wirral[18] made some laudable effort to reduce test data to indigenous surnames by utilising onomastic sources from the 15th century and the results showed no ancient M222 in this area of north west England. Another interesting study was reported by Busby and Myres in 2011 which typed M222 in the same Irish samples used several years earlier by Moore alongside samples from Britain and continental Europe[19]. They claimed an eyebrow-raising 44% of men in Ulster[20] were M222.

Another branch of research was a series of books by anthropologists, historians and geneticists trying a multi-disciplinary approach to the ethnic history of the Isles. These were sometimes hard to follow when the authors made up their own monikers for haplogroups. The 2011 book by Moffat and Wilson showed a map of M222 distribution in Ireland and Britain peaking in Ulster and Connaught and attributed it without reference to the direct progeny of Niall Noígíallach [21]. The most notable of these books was that by Mallory[22] in 2013 which was entirely based around Niall as the definitive first recorded Irishman. But despite using a multi-disciplinary approach examining upstream SNPs and archaeological and linguistic data Mallory fails to challenge the Moore et al. basic conclusion. The most recent book in this genre to date is that of Manco, in which the author dismisses the Niall theory and instead speculates that M222 is actually a somewhat older La Tene marker[23].

---

[17] 'Irish genetics and Celts', McEvoy, B and Bradley D, in: Celtic from the West, ed. Cunliffe B and Koch, J. (2010).
[18] King, Turi et al. Excavating Past Population Structures by Surname-Based Sampling: The Genetic Legacy of the Vikings in Northwest England. Molecular Biology and Evolution 25/2 (2008) 301-9.
[19] Busby G. et al, 'The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269', Proc. R. Soc. B. (2012) 279, 884-892
[20] Busby G. et al. op. cit., Data supplement (TableS1-HG frequencies) http://rspb.royalsocietypublishing.org/content/suppl/2011/08/18/rspb.2011.1044.DC1/rspb20111044supp2.xls, downloaded Dec 30th, 2013. The 'North Ireland (IRE-N)' region appears to derive from the 2002 Moore et al. Ulster sampling region.
[21] Moffat A & Wilson J., The Scots. A genetic Journey pp148-9 (2011)
[22] Mallory, J.P.. Origins of the Irish. (2013, London).
[23] Manco, J. Ancestral Journeys. The peopling of Europe from the first venturers to the Vikings, p189 (2013).

Despite this, the groundwork for the real breakthrough was already underway, partly by luck. Of the first ten humans to have their genomes fully sequenced, no less than three were M222 derived, led by Jay Flatley, the CEO of Illumina, Inc who was sequenced in May 2009[24] quickly followed by African-American father and son Henry Gates in Aug 2009 – numbers 6, 7 and 9 on the all-time list. Other samples appeared in projects like the Personal Genome Project and the 1000 Genomes project. Each year genetic genealogists talked excitedly about the so-called 'thousand dollar genome' but it always seemed to be coming next year.

Part five. Next generation sequencing and the splitting of M222.

Despite a false dawn when the Genographic 2.0 Project was rumoured to have several SNPs splitting M222 (they turned out to be near private and/or unstable), it was only towards the end of 2012 and into 2013 when next generation sequencing struggled into the affordable arena. Commercial R&D by ScotlandsDNA sequenced two Scottish M222 males and brought out a commercial SNP chip with all their approximately 27 SNPs. During the commercialisation phase of this data, another academic announced discovery of another significant branch SNP DF85 and this was added to the commercial chip. The chasing pack got their chance when a new startup FullGenomes Inc. offered pricey but just about affordable Y chromosome sequencing and a few months later a bigger rival came out with a more limited and cheaper option which still promised SNP discovery if not quite as comprehensively as FullGenomes (typically having one third to one half the chromosome coverage).

The impact of all this on the M222 research community was dramatic, going from many lean years making very little progress to suddenly being almost overwhelmed by testing options and new SNP names. At the time of writing (July 2014) there are in the region of 28 branching SNPs below M222 grouping people with different surnames, plus a small number of what might be considered 'private' SNPs linking people with the same surname[25].

Part six. Where next?

It is likely that work outside academia will continue to force the pace of research, albeit often hampered by a lack of neutral sampling. Data emerging from the UK-PGP project and the People of the British Isles project may help in this respect and the former at least should be releasing all its data into the public domain. The academics themselves seem to not be interested in revisiting the subject itself, eg Dr. Turi King[26] at the Leicester University Diasporas project, one of the foremost academic experts on combining data from Y chromosomes and surnames.

Postscript (April 2017)

The basis structure of the M222 tree has changed little since 2014. The three major level one branches remain S658, S568 and FGC4077 and almost everyone who is

[24] World Personal Genome Registry http://www.worldpgr.com/profile.aspx?ID=19 accessed Jan 2nd 2014.
[25] See http://www.kennedydna.com/M222_tree.png and http://www.kennedydna.com/M222.pdf

[26] King, Turi, Feb 2014 pers. comm

tested falls into one of those. A small fourth branch A7362 has been identified. Further down new branches continue to be found but for a complete perspective it is best to consult trees and a full description of them is outside the scope of this document.

There have been some minor adjustments to age calculations of both M222 itself and the sub-branches, with M222 currently estimated to have a TMRCA of 1900y (95% CI 1600-2200y).

Several groups work on the Y chromosome tree either in part or in full. Of those doing comprehensive Y trees, by far the best is that of 'YFull' who typically update their tree every month or so, complete with age estimates of each branch point. The full service that YFull provide for the token fee of USD49 is remarkable and includes both SNP and STR matching, age calculations, raw data browsing and private messaging with matches.

In addition YFull provide haplogroup (SNP) projects and the M222 project, open to YFull M222+ participants only, can be found at www.yfull.com/groups/R-M222.

Sequencing, which is needed to find new Y tree branches, has continued to become more mainstream whilst at the top end of the market advancing in technology for those who wish to be pioneers and can afford the best. Whilst the best tests remains those sold by Full Genomes Inc. of the USA, Dr. Thomas Krahn at YSEQ has also started to offer whole genome sequencing at his Berlin lab.

For those not wishing to make the commitment to sequencing there continues to be a good choice of fixed SNP tests. For those who know or strongly suspect they are M222 the YSEQ lab is very good. For those who have never done a Y chromosome test a bulk SNP chip may be a better bet.

One area in which M222 has drawn a blank so far is ancient DNA or aDNA. This is a rapidly growing field and many interesting discoveries have been made concerning Y DNA further back in time and higher up the tree than M222. The first aDNA from Ireland was extracted from Rathlin Island off the Co. Antrim coast and whilst this might have been M222 given its location it was reported in 2016 to be DF21. It should only be a matter of time before we have the first M222 aDNA, although it is doubtful this will end the arguments about country of origin.

Finally, on a personal note I was honoured to be selected for an award in late 2014 as 'Genetic Genealogist of the Year' for my work on the M222 tree. I shared the award with the well-known genetic genealogist CeCe Moore who was recognised for her autosomal DNA work. Most of my work more recently has been via the YFull M222 project which I co-admin with Aidan Byrne, or more focussed on the FGC4077 branch (as well as more general surname DNA work which covers non-M222 testers).