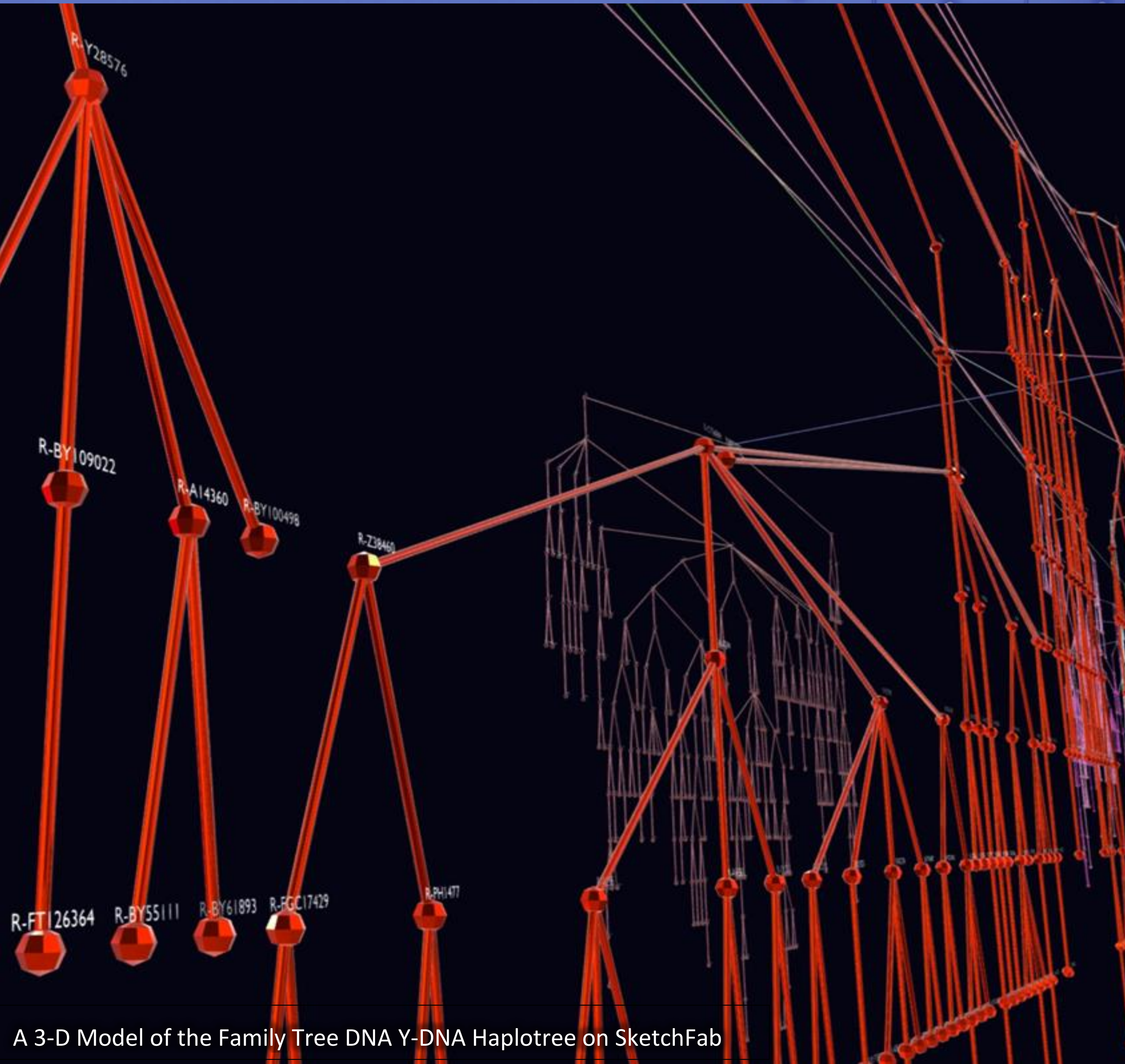
Journal of Genetic Genealogy

Volume 9 Number 1 (Fall 2021)



A 3-D Model of the Family Tree DNA Y-DNA Haplotree on SketchFab

Journal of Genetic Genealogy

Inside this Issue *(table is clickable in PDF Readers)***:**

https://www.jogg.info

Journal of Genetic Genealogy

EDITOR'S CORNER: JOGG RENEWED

By J. David Vance

Aaaaaand…we're back! With this new Journal of Genetic Genealogy Volume 9, Number 1 (Fall 2021) issue, the JoGG picks up again with a new issue and a new website.

The five years since our last issue have been maturing years for genetic genealogy. Commercial autosomal DNA databases now house well over 40 million tested members; Y-DNA testing has identified nearly two million named SNPs; archeological digs now routinely test DNA from human remains and investigative genetic genealogy has become a permanent tool in forensic analysis. And yet our testing technologies continue to evolve and our ability to use DNA to answer genealogical, anthropological, and forensic questions continues to evolve as well. As a discipline we are certainly not yet mature.

Much paper has been expended by those much more expert than I in describing how scientific and humanities disciplines evolve and change, and I won't try to say that genetic genealogy is in any particular stage of development or is even at any particular inflection point in our relatively short history. It's enough to point out that like most disciplines we do evolve, and that the Journal of Genetic Genealogy is here partly to document that evolution and partly as a resource for those new to the field to "catch them up" to how far we've come. But we don't want to be just a passive "best practices" repository – we also want our articles to illuminate the path of our forward evolution for the genetic genealogy community.

This last purpose is then also a challenge to the community: did we get it right? Evolution is not really linear so my metaphor of a "path" is flawed and our evolution is much more complex. What have we missed? What advances have not yet been documented, what methods need more recognition, what stories remain to be told?

The Journal of Genetic Genealogy is a forum for many stories: peer-reviewed scholarly articles, editorials, case studies, and introductions for new members of our community as well. We have examples of all of those in this new issue. An editorial invites us to consider whether a subfield of genetic genealogy should receive the additional focus that a differentiating name can confer. A master's dissertation originally presented to the University of Strathclyde in Glasgow in 2016 invites us to consider autosomal DNA analysis in new ways. We also have a review of network clustering visualization as a supporting autosomal DNA analysis tool, and another article provides some bridges between methodologies and practical applications of Time-to-Most-Recent-Common-Ancestor timeframe estimations using Y-DNA.

We are also introducing some regular feature series to focus on case studies and "how-to" methodology as a source of practical tips for administrators and other analysts. A feature with a Y-DNA focus starts off in this issue as a series called "Word to the Ys" with an explanation of short tandem repeat analysis through analogy. The mtDNA series will be called "mt Space" and we would welcome any mtDNA leaders to contribute. We also want to showcase your practical examples of genetic genealogy in action especially mixing different methods or different types of DNA testing, and the Miller case study in this issue is a great example of that. We plan other regular series on autosomal, mixed DNA analysis, or any other relevant subfield of genealogy as well; they haven't been named yet because I couldn't think of suitable puns so feel free to offer a feature series name suggestion with your first article!

But more importantly, please help us achieve our ongoing challenge to the genetic genealogy community by contributing your examples and critical thoughts on any subject relevant to genetic genealogy as articles for upcoming issues. You can find ideas and templates on our website at https://jogg.info. Or if we can help you translate an idea or case study into an article, please reach out to me at editor@jogg.info. You can help us document, you can help us explain and inform; but perhaps more importantly you can help us illuminate the path (however complex) of the forward evolution of genetic genealogy.

Our new website look also includes an archive of all our past journal articles indexed with article categories and tags so you can more easily find interesting material in our back issues. We hope you find this archive a useful resource in your personal evolution in genetic genealogy as well.

Journal of Genetic Genealogy

Would GENEALOMICS be an appropriate term to designate family tree research based on genome-wide data?

***Alho**, Clarice Sampaio*[1,2]; ***Dorn**, Marcio*[2,3]; ***Avila**, Eduardo*[1,2,4,*]

[1] Laboratory of Human and Forensic Genetics, School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil.
[2] National Institute of Science and Technology – Forensic Sciences, Porto Alegre, Brazil.
[3] Laboratory of Structural Bioinformatics and Computational Biology, Informatics Department, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.
[4] Technical-Scientific Section, Rio Grande do Sul State Regional Division, Brazilian Federal Police, Porto Alegre, Brazil.
[*] Corresponding author: e.avila@edu.pucrs.br

ABSTRACT:

Current abundance of available human individual genomic data has allowed the advent of genealogy applications based on DNA information. Family trees and distant kinship relations can be assessed using new proposed genetic data and analyses methods. This work proposes and discusses the feasibility and pertinence of employing the term "*Genealomics*" as a comprehensive term including genetic data and processing methods focused on DNA-based genealogy applications. More than just a trendy adoption of another "*-omics*" derived term, this concept is useful to delineate the range of genomic information and technical procedures that can be used with genealogic purposes, including in a forensic context.

KEYWORDS:

genealogy; genetics; investigative genetic genealogy; kinship; genealomics

Genealogy is understood as the research based in documental (primary or secondary) sources to demonstrate kinship and relationships among members in a family tree, pedigree, or lineage. Genealogy research is commonly based on historical reports, official documents and pictures, biographies, civil or military records regarding life events (birth, marriage, and death certificates, for instance), verbal statements, judicial records (as inventories, testaments, and other legal contracts), historic or biographical reports published in old newspapers and almanacs (personal announcements and other public proclamations). Obituaries also comprise a rich source of information, since they usually contain several biographical facts about the deceased and family or relatives, medical records and other sources.

Current genealogical research focuses not only on reconstruction of family trees or biogeographical ancestry determination due to personal interest or curiosity. Instead, such studies are also performed following court request, considering its legal implications, via Forensic Genealogy. Examples of genealogy applications in a forensic context are very common, including efforts to establish, grant or determine citizenship status of an applicant, admitting relatives or finding heirs, or demonstrate close or distant kinship relationships. Since most legal systems ensure inheritance rights to children or other relatives after someone dies, applications in civil or family law are largely widespread. Application of genetic and molecular biology tools to assist in this investigation is also a current practice: genetic analyses of biological samples are extensively employed to determine kinship in court cases, to determine paternity or other biological relationship between individuals. Paternity determination, identification of human remains and disasters victims, investigation of missing persons, human trafficking and illegal adoption practices, and a wide range of other types of criminal activities have been investigated using forensic genetics methods and techniques.

When compared to traditional genealogists, geneticists, correspondingly, build, propose, compare or confirm pedigrees through Genetic Genealogy, essentially based on DNA data, to recognize the genetic inheritance from a particular parent when a child is born. Law enforcement has applied the term *Investigative Genetic Genealogy*, or *IGG*, when DNA profiles from a crime scene or from unidentified human remains are used for human identification purposes by comparing the known genealogy of possible close relatives [1]; IGG is also known as *Forensic Genetic Genealogy*, to describe DNA-based relative combined matching.

In biosciences, the suffix "*-ome*" (from the Greek –ωμα) is applied to form nouns in the sense of referring to a complete whole of a class of substances or data for a species or an individual, in which their constituents and interrelations are collectively and simultaneously considered [2-3]. Such definition usually

includes datasets comprising substantial volume of information [4]. Many "*omes*", from the original Genome, first coined in 1920 [5-6], have been adopted by scientists as a term for large-scale analyses and connections in a specific field [7]. From the first sequenced human genome with over three billion base-pairs, current technologic advances led to thousands of complete genome data, which are now widely available. This number is increasing proportionally to sequencing and bioinformatics technology improvements, consolidating the new field called Genomics in order to distinguish and typify organisms, structures or systems from a wide genome data. The radical "*-omics*" is a derived neologism applied to studies that collectively identify, map and characterize big-data of biological molecules involved in structures, profiles, pathways, and composition of cells, tissues, organisms, species, or populations.

Expansion of massively parallel DNA sequencing facilities have provided large volumes of genomic investigations. A vast amount of genomic information from volunteers, obtained from publicly available research datasets or tested in Direct to Consumer-DTC services, is now accessible. Genealogical exploratory approaches using the genomic information from these collaborators have been the means to investigate kinship links among individuals. Investigators and researchers have been able to successfully identify consanguinity connections, applying methods to recognize identical-by-descent (IBD) DNA segments. Such approaches have been adopted to detect familial matches sharing DNA segments, even when the biological relationship is relatively distant [1]. In the same way as Phylogenomics studies procedures and techniques designed to propose fully resolved phylogenetic trees, methods employing exhaustive analyzes in large-scale databases with simultaneous comparison of large numbers of highly dense or full-length human genetic profiles, may be required to allow recovering of genealogical relatedness correspondences [8].

Family search methods based on genome-wide big data require both the massive amount of information contained in full length DNA sequences databases and several specific-tailored computational approaches to precisely calculate the degree of relatedness between the contributors. As it intersects methods from the larger fields of genealogy, genomics and bioinformatics, we introduce the term *Genealomics* to refer to such strategy. *Genealomics* is an interdisciplinary field that draws information by computationally comparing entire genomes or large genetic variant datasets (containing thousands to millions DNA polymorphisms) to establish and clarify their kinship or common ancestry relationships; the term could be applied in multiple ways to represent any analysis involving genomic full-data or large-scale microarray genotype data and family reconstruction informatics procedures to predict or suggest personal biological interconnections. Although not exhaustive, the list of potential applications of *Genealomics*-based methods includes the mapping and identification of endogamy levels and pedigree collapse, which can be estimated for a single individual or for a group of people belonging to a local, regional population. Other applications might include reconstruction of historical genomes, belonging to early settlers of specific areas or other people of interest, by combining IBD fragments of a large number of relatively distant descendants. Such applications, of course, would depend on study and availability of across-population level genetic datasets, which might be hard to obtain or have limited access due to ethical or privacy concerns.

Coining a new term, often consisting of a combination of other previously defined words, is reasonable in terminology definitions when it emerges from an accepted, coherent, and well-known logic; we assume the *Genealomics* definition can be useful in optimizing and promoting a supportive keyword for genomic genealogists. Considering it comprises a subfield of Genetic Genealogy, and a particularly technical one, adoption of specific language can successfully drive conceptual structure and compartmentalization. The endorsement of distinguish jargon and terminology by the Genealogy community can serve to highlight a particular subdiscipline deserving of independent focus to establish methods and practices either unique to or adapted for that subfield. Adoption of this concept and terminology may also include the analysis and evaluation of documental (as provided by traditional, not DNA-related genealogical methods) support for kinship or biological relatedness. This consideration is especially relevant when considering the construction of genomic reference databases for a particular population, when clustering methods can be employed to narrow down possible family lines related to a particular sample. Legal and ethical concerns regarding this particular approach must also be discussed in order to verify how the *Genealomics* concept can be adopted in law enforcement and legal cases. Even though some criticisms have been published about the extensive use of "*-omics*"-derived terms, it is acknowledged that specific nomenclature regarding subsets of genomic applications can be useful in some cases [9-10]. The specificity of analytical methods and genetic information employed in DNA-based genealogical approaches may be considered sufficient to justify the adoption of a particular definition. A concise, direct terminology might help a deeper understanding and broader dissemination of technical concepts associated with these applications to scientific and legal communities or even to the general public, as traditional or genetic genealogists and clients from DTC companies.

Furthermore, it is important to mention that traditional DNA-based methods for kinship determination ca be highly improved by *Genealomic* analyses, where both a pipeline of high computational complexity and very dense panels of DNA markers may lead, in particular cases (especially those including distant biological relatedness or where traditional techniques are insufficient), to more precise and accurate relationships evaluations; studies propose that a large number of nucleotide variations are necessary to determine siblings (at least tens of thousands polymorphisms), while most distant biologic kinship relations would require increasingly higher genetic variants genotyping [8]. Likewise, to return few people with high IBD, e.g., two or more individuals sharing over 100cM (centimorgans) DNA segments inherited in haplotype phase without recombination from a common ancestor, a long-range familial search has to be traced within databases containing full-genomic or highly dense records from a large number of samples, preferentially containing thousands of individuals [1].

Genealomics is a proposed *omics*-inspired term, motivated by the new genealogical proceedings relying on DNA sequence comparisons to infer relationships, where numerous and extensive computational tools are required. It differs from traditional kinship-resolving methods, since it is based on large amounts of genetic information, in contrast to a relatively smaller number of highly polymorphic genetic variants (as short tandem repeats). As such, it has the ability to propose or identify more distant biological relatedness. Therefore, to the scientific community and other interested parties, we suggested the term *Genealomics* could be broadly used to refer to approaches aiming at personal interconnection and kinship prediction or estimation, especially those employing large-scale DNA data and including comprehensive population databases. It consists basically of a subset of the complete genomic data (evaluated at both individual and population levels), particularly the genetic information necessary or useful to investigate or determine biological relatedness among distinct individuals. The concept also includes all analytical methods, approaches and computational systems used to handle and process the biological data. By comparing vast numbers of highly dense DNA sequences or entire genomes among many people, *Genealomics* can be used to propose kinship relationships and identify how closely related they are. It is also important to mention that, as a recent advance derived from the exponential growth in available genomic data, as well as modern progresses provided by bioinformatics research, the need for extensive research, development and improvement in both field-related methodologies and proper use (considering legal, ethical, and technical aspects of available personal genome-wide data) is essential. The contribution of *Genealomics*-derived methods to traditional and genetic genealogy has the potential to significantly improve family and pedigree studies, with implications on many fields as kinship mapping or forensic investigation. The genealogy community is challenged to explore the possibilities presented by this promising area.

ACKNOWLDGEMENTS:

The authors would like to thank the reviewers and editors for their insightful contributions and suggestions to the present work.

[1] Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: Current methods, knowledge and practice. Forensic Sci Int: Genetics. 2021;52: 102474. doi: 10.1016/j.fsigen.2021.102474

[2] Yadav SP. The Wholeness in Suffix -omics, -omes, and the Word Om. J Biomol Tech. 2007; 18(5): 277 PMCID: PMC2392988 PMID: 18166670

[3] Kuska B. Beer, Bethesda, and biology: how "genomics" came into being. J Natl Cancer Inst. 1998;90(2):93. doi: 10.1093/jnci/90.2.93.

[4] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020; 14: 1177932219899051. doi: 10.1177/1177932219899051

[5] Lederberg J, McCray AT. `Ome Sweet `Omics - A Genealogical Treasury of Words. The Scientist. 2001;15(7): 8. Gale Academic OneFile.

[6] Goldman AD, Landweber LF. What Is a Genome? PLoS Genet. 2016;12(7): e1006181. doi: 10.1371/journal.pgen.1006181

[7] Myers AJ. The age of the "ome": genome, transcriptome and proteome data set collection and analysis. Brain Res Bull. 2012;88(4):294-301. doi: 10.1016/j.brainresbull.2011.11.015

[8] Erlich Y, Shor T, Pe'er O, Carmi S. Identity inference of genomic data using long-range familial searches. Science. 2018;362: 690–694. doi: 10.1126/science.aau4832

[9] Petsko GA. No place like Ome. Genome Biol. 2002; 3: comment1010.1. doi: 10.1186/gb-2002-3-7- comment1010.

[10] Eisen, JA. Badomics words and the power and peril of the ome-meme. GigaScience. 2012;1(1):6. doi: 10.1186/2047-217X-1-6.

Journal of Genetic Genealogy

WORD TO THE Ys: UNDERSTANDING Y-DNA STR GROUP ANALYSIS THROUGH ANALOGY

By J. David Vance

“Word to the Ys” is a new feature series where guest authors write about experiences and methods for using data from Y-DNA testing in the practice of genetic genealogy.

Introduction

I apologize in advance to those who don’t like analogies, but I find them useful; and especially powerful when people are struggling with a particular concept and the analogy forces them to think about the concept in a different way that helps them makes sense of it. Analogies are never perfect of course, so you have to take them in context of whatever point they were being used to make and know that they are probably not otherwise applicable.

This analogy is really for people who may be still struggling with STR allele values as numbers and trying to understand how they can be used to analyze a subgroup of a surname project (the same techniques can be used with subgroups of haplogroup projects or other types of projects, but it’s easier to explain using a surname project).

Please understand that this is only a simple introduction and the analogy is by no means perfect. It certainly offers no insight into the biology of DNA, it glosses over some of the complexities of STR analysis, and it does not address SNPs (like you get from Big Y700 testing or Whole Genome testing) at all.

Also I’m just explaining a way to use STRs, not advocating that using them alone is the best strategy. In fact I usually advocate for using both SNPs and STRs if at all possible. Many people want to jump straight into SNPs and ignore STRs completely, and while there is no doubt that SNPs can give structure and more certainty to a group analysis if you can include them, there are many examples where STR analysis can be useful, like:

- If you’ve taken on the task of analyzing a group that has already extensively tested STRs, and you want to make best use of the investment they’ve already made in testing before asking them to upgrade to Big Y700 or other SNP tests;

- If your group doesn’t have the available funds to do extensive SNP testing;

- Or if the group has already done extensive SNP testing but key questions still remain about the branching within the group because no SNPs have yet been discovered that mark those branches. STRs can often be used to find additional branching in between SNPs.

But hey, if you’d prefer to ignore STRs completely and only work with SNPs then that’s fine and this analogy is not for you. Feel free to skip it.

Before we start out with a group analysis it's also important to know what we're trying to find out. In this analogy we have a surname project subgroup that wants to know how they are all connected to each other; this is a common question but by no means the only one. Are they trying to break through specific brick walls? Are they trying to connect back to a particular ancestor of some importance? Are they trying to show a lineage to a particular ancestral group like a clan or specific community? Are they trying to find a common region of origin? These will all guide your purpose in analyzing the group and what information is important to look at, because it's probably why they invested in Y-DNA testing in the first place. Of course the other possibility is that you paid for their testing for a specific purpose, in which case that's the purpose that's important. So again in this example we'll be trying to build a genetic family tree that connects a group that don't already know how they're connected, but take that as just one example among many and tailor your approach to what your group is interested in finding out.

Ok so with that preamble over, here's the analogy.

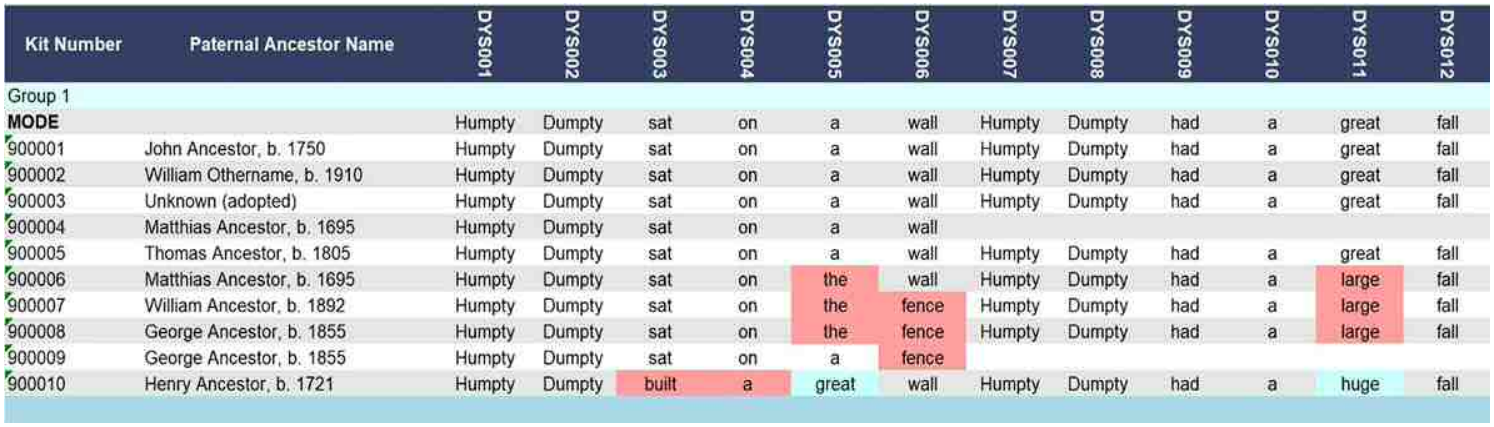
The Analogy: STRs as words in a Nursery Rhyme.

Suppose a man long, long ago taught the nursery rhyme "Humpty Dumpty" to his sons, and started a tradition in which every one of his male descendants taught that same nursery rhyme to his own sons in turn. The tradition has never been broken, but occasionally a father mixes up a word here and there in the rhyme when teaching it to one of his sons, so the rhyme can change slightly as it is passed down.

So this first ancient ancestor taught his sons "Humpty Dumpty sat on a wall, Humpty Dumpty had a great fall, all the king's horses and all the king's men couldn't put Humpty together again." And his sons taught their sons the same rhyme, but maybe one of his sons mis-remembered and said "Humpty Dumpty sat on the wall" and so on, and his own sons taught it that way to their sons, and over time the rhyme has been slightly altered on certain male descendant lines from this original man.

As we'll see in a minute, the other important part of the analogy is that some words in the rhyme are more easily changed than others. "Wall" and "Fall", for instance, are easy to remember because they sound the same and because they're important to the meaning of the rhyme, so you wouldn't think a man would change those words very often. On the other hand, a determiner like "a" could easily be changed and you could easily see how instead of "a wall" a man could teach his son "the wall" or even "some wall". In our analogy, words that don't change often will take the place of slow-mutating markers, and words that can more easily change will take the place of fast-mutating markers (and I'll note that often it's not really important what the exact mutation rates are; for analysis purposes it may just be important to see if you're working with particularly fast-mutating or slow-mutating markers).

It's now present day, and you have gathered together 10 of the descendants of that original man, who have all done some level of genealogical research but they don't know how they're all connected. You test them each to see what they remember of the start of the nursery rhyme, and 8 men take a "12 word/marker" test and 2 men take only a "6 word/marker" test (for simplicity, I've used fewer "markers" than actual Y-DNA STR tests, but you get the idea). You put all the results in the following table. For ease of reference, we'll label these first 12 words and we'll call them "DYS001" through "DYS012" so we can refer to them later.



Kit Number	Paternal Ancestor Name	DYS001	DYS002	DYS003	DYS004	DYS005	DYS006	DYS007	DYS008	DYS009	DYS010	DYS011	DYS012
Group 1													
MODE		Humpty	Dumpty	sat	on	a	wall	Humpty	Dumpty	had	a	great	fall
900001	John Ancestor, b. 1750	Humpty	Dumpty	sat	on	a	wall	Humpty	Dumpty	had	a	great	fall
900002	William Othername, b. 1910	Humpty	Dumpty	sat	on	a	wall	Humpty	Dumpty	had	a	great	fall
900003	Unknown (adopted)	Humpty	Dumpty	sat	on	a	wall	Humpty	Dumpty	had	a	great	fall
900004	Matthias Ancestor, b. 1695	Humpty	Dumpty	sat	on	a	wall						
900005	Thomas Ancestor, b. 1805	Humpty	Dumpty	sat	on	a	wall	Humpty	Dumpty	had	a	great	fall
900006	Matthias Ancestor, b. 1695	Humpty	Dumpty	sat	on	the	wall	Humpty	Dumpty	had	a	large	fall
900007	William Ancestor, b. 1892	Humpty	Dumpty	sat	on	the	fence	Humpty	Dumpty	had	a	large	fall
900008	George Ancestor, b. 1855	Humpty	Dumpty	sat	on	the	fence	Humpty	Dumpty	had	a	large	fall
900009	George Ancestor, b. 1855	Humpty	Dumpty	sat	on	a	fence						
900010	Henry Ancestor, b. 1721	Humpty	Dumpty	built	a	great	wall	Humpty	Dumpty	had	a	huge	fall

Figure 1. The group of descendants with up to 12 markers tested.

(and yes, this picture is formatted on purpose like a project's DNA Results page just to make the analogy clearer that we're using words in place of the STR numbers that you'd usually see).

From a regular genealogy standpoint, two of the men have a different surname, but one of the two was adopted anyway and doesn't know his biological father. The rest of them all carry the same surname ("Ancestor"), and several have traced their ancestry back to the same earliest ancestors but those are all in different centuries and while it's clear that they're probably all related, we don't yet know how.

In our picture above, we only see the earliest known paternal ancestors for each of the kits, but as a project administrator it's important also to collect the whole known paternal lineage from each member including the regions where their ancestors lived and any information, even hints, that they might have about their male paternal line. I'll give a few examples in the rest of this introduction about how this information can help us in building our genetic family tree.

So now let's stare at the "markers" for a bit.

The "Mode" row, also called the "modal haplotype" or just "modal", is the collection of the most commonly recurring words among the group, and clearly "Humpty Dumpty sat on a wall, Humpty Dumpty had a great fall" is the "modal haplotype" for this group. Because words don't change very often, the most common set of words is likely to be the set of words that the common ancestor of this group first taught his sons, which is known as the "ancestral haplotype". It's not guaranteed that the modal haplotype and the ancestral haplotype are always the same, but it's usually a good starting assumption especially if you have more than a few testers, and normally if the modal and ancestral are different it's only on one or two markers. In this introduction because we're keeping it simple, the Mode will be the same as what the common ancestor taught his sons. Of course real STRs also have a Min and a Max row which can occasionally come in handy, but since words don't have a "min" or "max" I've left those out here.

So then we have "off-modals", meaning words that are different from the Mode values. These are highlighted in red and blue in the earlier picture. Since STRs are numbers of course those can go up and down; in this analogy we just have words that have "mutated" to other words.

So how do we put this puzzle together? Can we tell how these men are related to each other?

Looking for Strong STR Patterns

Let's start with kits 90007, 90008, and 90009, because we can immediately see that they have a lot of "off-modals" in common. They all have "fence" as an off-modal for marker DYS006, and the two who have tested out to 12 markers have "large" for DYS011 instead of the modal "great". Changing "wall" to "fence" is not an easy mistake and it's not likely to happen very often (an analogy for a very slow-mutating STR marker), so it's very likely that only one father some time in the past said "fence", and his sons passed that mutation of the rhyme on to their own descendants.

It's very tempting to throw 90006 into this analysis immediately also because he shares the off-modals "the" for DYS005 and "large" for DYS011 with kits 90007 and 90008. But he doesn't have "fence" for DYS006, and 90006 also apparently shares the earlier Mathias Ancestor b. 1695 as an ancestor along with kit 90004, who doesn't have "the" for DYS005. So DYS006 is likely closely related to this 90007/90008/90009 subgroup, but let's look at the smaller subgroup of those three first and then we'll widen our search (this is a common strategy by the way; look at small groups that share key markers first and then expand the analysis out to include other kits from there).

We don't know for sure that kit 90009 also has "large" for DYS011 since he hasn't tested it, but after some study we can conclude that it's very likely that he does, both because everyone else who has the off-modal "fence" for DYS006 also has the off-modal "large" for DYS011 and also because kit 90009 has traced his ancestry back to George Ancestor b. 1855 just like kit 90008. This is one example of where it's important for you to collect the known genealogy information for the group so that you can use it to guide your conclusions. For this analogy of course we don't know for sure that their traditional genealogy research is correct, but let's assume you've checked and decided it's reliable and you can safely conclude that the chances are very high that kit 90009 also has "large" for DYS011 at least until you upgrade him to more markers and confirm it.

So kits 90007, 90008, and 90009 all share the "off-modals" "fence" for DYS006 and "large" for DYS011. Both are important words to the meaning of this nursery rhyme and not easy mistakes to make (our analogy for slower-mutating markers) and so a father changing either on its own would be unusual, but the combination of both in this subgroup of 3 kits makes it statistically near-certain that they share a more recent common ancestor with each other than with the rest of this group. When two or more slower-mutating markers like this point to a subgroup of a larger group, this is a recognizable pattern that says these kits are descended together from one or more common ancestors who (perhaps in one man or perhaps over a series of generations) developed this pattern. The stronger this pattern (the more markers that form it, and the slower-mutating they are), the more certainty this provides about the branching.

In many larger haplogroups, experienced project administrators have created what are called “allele frequency tables” which can tell you how rare a particular STR value is within that haplogroup. These can be useful to assess the strength of a particular STR pattern, because rare marker values are especially more likely to indicate common branches. So for example, kits 90007, 90008, and 90009 all have the value “fence” for marker DYS006. That word destroys the rhyming scheme of our “Humpty Dumpty” rhyme since it doesn’t rhyme with “fall”, and so it would be particularly unlikely for a man to teach the rhyme to his son with “fence” for DYS006. With our analogy we could say that “fence” is an example of a rare marker value and so a pattern that includes it, especially in combination with other unusual markers, could be considered particularly strong.

Patterns like this are sometimes called a “STR signature” that marks a subgroup, or just a “unique STR pattern”, or especially for European or British Isles English speakers, it is also sometimes called a “STR motif”. The more markers there are in the pattern and the slower their mutation rate, the more likely it is that they mark a subgroup that sits on its own branch, and very strong “signatures” are just as good as SNPs for that purpose.

Now we look at marker DYS005 and notice that both 90007 and 90008 have the off-modal “the” instead of the modal “a”. Marker DYS005 though is a determiner in the sentence and not a very important word (our analogy for a faster-mutating marker) and certainly it would not be hard for a father to mis-remember that word so we could expect it would easily change – on its own it’s not even close to a strong pattern, but since we know the genealogy for both 90007 and 90008 is reliable and they work back to a common ancestor we can conclude that one ancestor in their common line (either George Ancestor or one of his immediate patriline ancestors) had the mutation “the” from “a”. If we diagrammed this situation, it might look something like this:

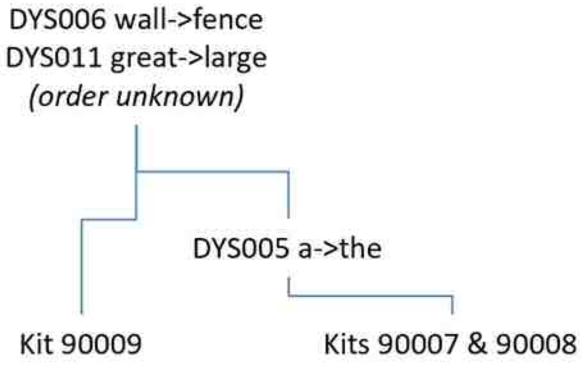


Figure 2. A first mutation history diagram

In the diagram the mutations are marked where they occurred in the ancestral lines, so first one of the common ancestors of all three kits changed “wall” to “fence” and “great” to “large”, and then somewhere after kit 90009 split off, one of the ancestors of kits 90007 and 90008 changed “a” to “the”.

Notice by the way that we don’t know which of the DYS006 and DYS011 mutations happened first; if the same ancestor changed the words for both DYS006 and DYS011 or if they were changed by different men, we only know they both happened somewhere in that range of ancestors. If we did have two or more descendants tested from every generation in this male ancestry then we could easily tell the order of these mutations, but that’s not a realistic expectation. In practice we can only assign certain mutations to certain ranges of ancestors, and the order in which those mutations occurred is unknown (and exactly the same thing occurs with SNPs by the way which leads to equivalent blocks of SNPs).

Next: Finding the most likely scenario among several possibilities

Now let's look next at kit 90006, because he shares the "large" mutation for DYS011 and the "the" mutation for DYS005, but NOT the "fence" mutation for DYS006. So he doesn't immediately fit in the branches of our picture above. In fact when we try to include 90006 in our diagram there are three main possibilities for how his version of the rhyme might have come about:

1. Maybe he doesn't fit anywhere in these branches and he's a descendant of a separate line who happened to also make the same two mistakes and also taught their children "large" for DYS011 and "the" for DYS005. This is possible, but statistically not very likely because it means both DYS011 and DYS005 had to have two separate parallel mutations on different branches.

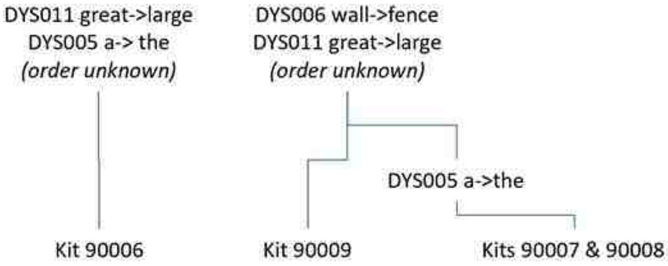


Figure 3. Possibility 1

2. Another possibility is that he is a descendant of the same common ancestors, but one of his ancestors changed DYS006 back to "wall". If that happened, the most likely scenario is that he inherited the "the" word change on DYS005 from a common ancestor with kits 90007 and 90008, and his DYS006 changed to "wall" from there.

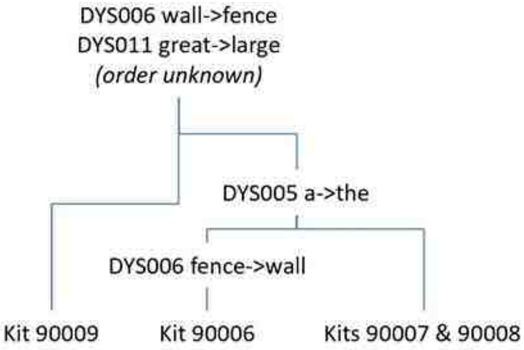


Figure 4. Possibility 2

3. The third and final possibility is that 90006 is related to this group still but earlier before the man who had the "fence" mutation on DYS006. That would mean we've guessed wrong on the DYS005 mutation; that it changed from "a" to "the" first earlier, and it changed back to "a" on kit 90009's line. That still leaves us three kits with "the" for DYS005 and one with "a", but it changes where DYS005 changed among these branches.

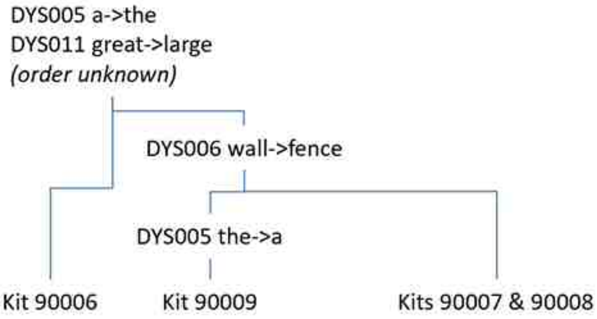


Figure 5. Possibility 3

In our first look at just kits 90007, 90008, and 90009, we found an example of a STR "signature", where two STRs were clearly pointing to a subgroup of men who were more closely related. Now we've moved beyond "clearly" into probabilities – we have several options that are each possible and we have to narrow them down to the most likely.

This is where many people get discouraged with STRs because they often only give you likely scenarios, not always certain ones. SNPs also actually have a range of likelihoods and don't always give you complete certainty, but on average their reliability is higher in supporting genealogy scenarios. If I were to diagram the relative reliability ranges of our various sources of information including traditional genealogy research, they might look something like this:

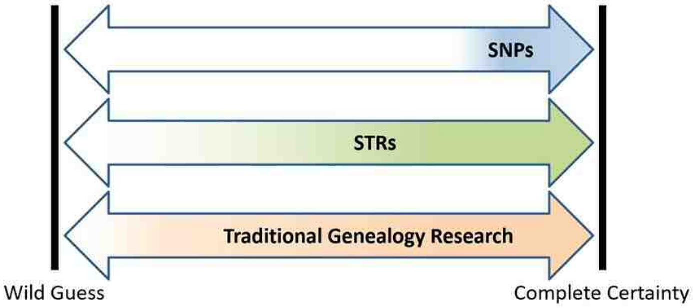


Figure 6. Subjective range of source reliability

With STRs, you can get as close to certainty as SNPs if you find really reliable patterns, but you do have a wider range of reliability in supporting various scenarios that you have to assess. This is not really different from traditional genealogy research, where we've always had primary and secondary sources that differ in reliability and have often found that some evidence (like the recollections of the past generations for example) is not always as reliable as we would have liked. So STRs and traditional genealogy research are very similar in that we have to assess them both for reliability, more so than with SNPs.

As I said earlier, perhaps the best approach is to combine all of these sources to answer your group's questions. In this introduction I'm focusing on STRs and talking a little about integrating information from traditional genealogy research, but if you do have SNP information from multiple SNP tests (like Big Y700 or Whole Genome testing) among the group this is even better, because SNPs may be able to give you some key sub-divisions that already break the group into clear subgroups even before you start analyzing the STR patterns. Depending on how many SNPs you can find that separate a group, you often can get the high-level skeleton of branches within the group from SNPs and then use STRs supplemented by traditional genealogy knowledge to fill in more detailed branching.

So turning back to our scenario, what can we use to distinguish between these three possibilities?

Well first of all, we notice that each possibility involves markers mutating more than once. In Case #1, DYS011 and DYS005 both had to change twice for a total of 4 mutations; once each on 90006's separate branch, and once each in the branches that 90007, 90008, and 90009 have in common. In Case #2, only DYS006 had to change twice, and in Case #3, only DYS005 had to change twice. The other thing that we need to remember is that we already said that DYS006 and DYS011 were slow-mutating markers, and DYS005 was a faster marker (using our analogy of the words that are less likely to be changed as being "slower-mutating' than words that could be changed more easily).

Statistically speaking, the scenario that involves the fewest number of mutations is most likely, and markers that mutate at faster rates are more likely to change than markers that mutate at slower rates. Following that logic, Case #1 is the least likely because it results in four extra mutations rather than two. And since we said DYS006 was a slow marker and DYS005 was a faster one, Case #3 is therefore statistically most likely, because only the faster marker DYS005 has multiple mutations.

When phylogeneticists (scientists who build evolutionary trees) build trees you will often see them applying a principle called "maximum parsimony", which essentially means trying to minimize the number of changes across the tree. If changes are already unlikely to begin with, then it stands to reason that the fewest number of changes (and additionally, the fewest number of less-likely changes) will result in the statistically most likely tree. But it's important to also remember that human ancestry doesn't always follow the statistically most likely path, so if this is the only evidence that we have to go on we can only conclude so far that Case #3 is suggested, but not that it's true.

The traditional genealogy knowledge might help strengthen either Case #2 or Case #3 also. Kit 90006 knows his genealogy much farther back than any of kits 90007, 90008, and 90009 (and let's say their earliest known ancestors aren't in 90006's line, otherwise they would also know they descended from Matthias Ancestor). That might argue for Case #3 also since 90006's connection to the others is much further back, but we don't know enough about that from this example. Perhaps also the regions these ancestors originated from or when their lines immigrated to other countries, etc, would make one of Case #2 or Case #3 either impossible or less likely.

It is also hard to know how much weight to put on the knowledge that both 90004 and 90006 list their earliest ancestor as Mathias Ancestor b. 1695. A lot depends on how reliable the genealogy research is for those two descendants, but it also matters how Mathias may fit into the genealogies of the other earliest known ancestors of this group. Could Mathias be a direct paternal ancestor of William Ancestor b. 1892 or George Ancestor b. 1855? Or does the traditional genealogy knowledge rule either or both of those out completely? Traditional genealogy research of course as we showed earlier also has a wide range of accuracy from "Wild Guess" to "Complete Certainty", but the details learned from traditional genealogy research may completely rule out one or more of these scenarios or they may only make one or more scenarios more or less likely. As the analyst of this group it's up to you to weigh the traditional genealogy evidence and decide that as well.

So perhaps we can say that Case #3 is only statistically more likely, or perhaps more evidence from traditional genealogy agrees or refutes that. For the sake of this example, let's go with Case #3.

Next: Kits that we can't yet place

Kit 90004 also traces his ancestry back to Matthias Ancestor b. 1695 like kit 90006, so it's tempting to add him into our Case #3 tree on the same branch as 90006. Notice that he has "a" for DYS005, so we might have to assume yet another mutation of DYS005 somewhere on our tree.

If we only made a decision based on the rhymes however, there's no visible reason to put 90004 and 90006 together. It would really be good to know the status of DYS011 for kit 90004 and if he has the signature off-modal "large" for that word, but he hasn't tested that far. There is no reason off-hand to believe his traditional genealogy research was wrong, but Y-DNA isn't giving us any real clues either way. If 90004's traditional genealogy knowledge was rock-solid and we also knew something about Mathias Ancestor's relationship to William Ancestor and George Ancestor then maybe we could draw some conclusions from there about how 90004 fits with 90006, 90007, 90008, and 90009 and maybe we could also place 90004 on our little diagram from Case #3 at least as a speculative or even possible branch. For the time being however, let's say we don't have enough reliable information to draw that conclusion.

When we look at kits 90001, 90002, 90003, and 90005, we see we know even less about them. These kits all have the exact modal values for all 12 markers – they received the rhymes unchanged from the group's common ancestor. That might mean that they all belong together on one branch, but it also might mean they belong on different branches that all had no mutations – we don't know. One thing we do know though is that they're not closely related to kits 90006, 90007, 90008, and 90009, since they don't share any of the same "off-modals."

Next: The out-of-place kit

Kit 90010 has a number of "off-modals" that no-one else shares. His ancestors taught him "Humpty Dumpty built a great wall, Humpty Dumpty had a huge fall." All four of those word variations in the rhyme are unique to his line.

In theory it's remotely possible that his ancestry does share one or two of those word variations with other kits in this group, but after 90010 split off their line "back-mutated" on those words back to the modal values. Without any reliable traditional genealogy knowledge that suggests that though, it's not statistically likely. Leaving our analogy for a minute, with STR numbers of course it's possible that mutations occur in steps in the same direction, so if the modal value is 12 and there is a sub-group that has 13 and another single kit has 14, you may have to consider that the kit that has 14 is a further mutation off the group that has 13. With words we can't illustrate that so we'll leave it out of this example.

Without any other evidence then, all we can conclude is that kit 90010 is most likely descended from an earlier branch of this group.

Putting it all together

So we have a combination of reliable marker patterns and possible scenarios, and if we threw it all together into one diagram it would look something like this:

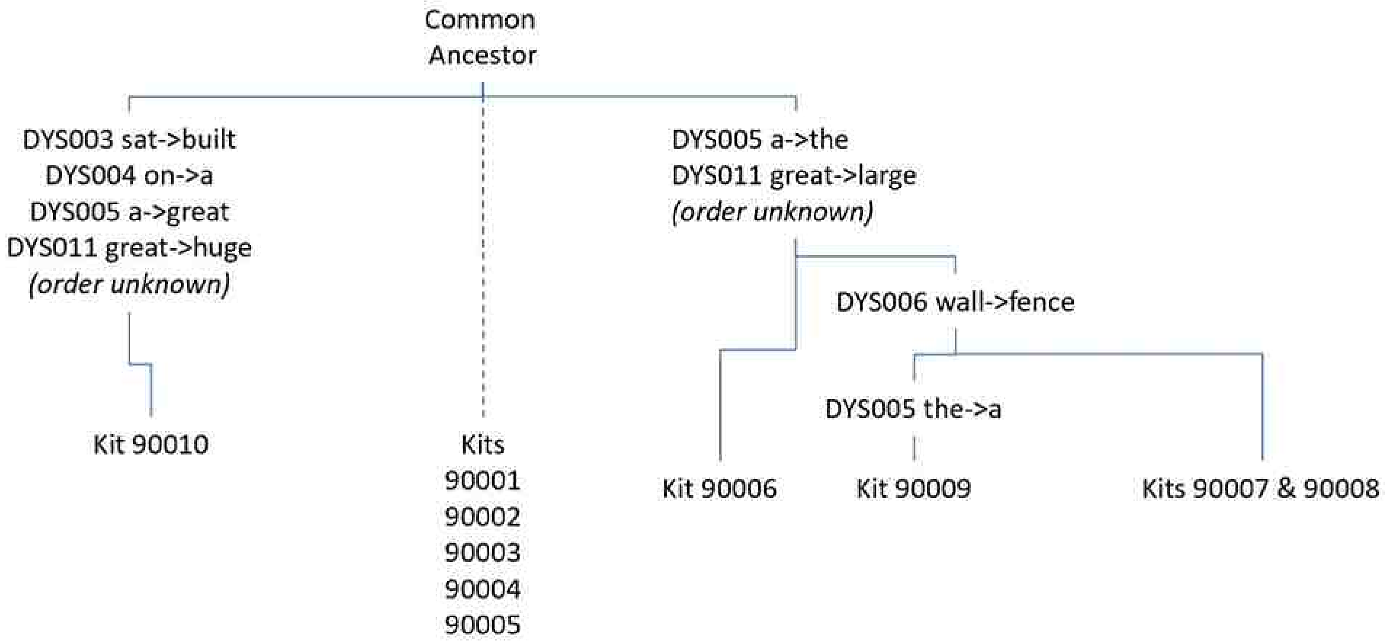


Figure 7. Starting to map the mutations on a tree

If kit 90004's traditional genealogy really was rock-solid of course then we might be able to list him over with 90006 and make an educated guess about how their lineage connected with the ancestry of kits 90007, 90008, and 90009. Since earlier we said we weren't sure about 90004's ancestry we've left that out of this "final" picture, but it's one variation that we might adjust depending on how strong we think that educated guess would be.

This type of tree diagram along with the associated mutation history is often referred to as a "mutation history tree" or "genetic family tree", and at the end of this introduction I'll recommend some resources for additional information and tips.

We can also of course start to add ancestors on the various branches based on what we know from traditional genealogy.

Unfortunately in this example we haven't learned anything about the likely NPE 90002 or the adoptee 90003. Both are probably related to this Ancestor surname group, but this level of information has not given us any clues as to where.

Next Steps

So where do we go from here? Obviously with this fairly simple analogy and small group of testers, we haven't discovered very much yet about their shared ancestry. But the example is still not unusual in that often you can draw some firm conclusions about a subgroup, that you have possible connections between others in the group, and that for some subset of the group, the DNA hasn't yet helped at all.

We haven't really spent much time in this introduction either on what additional traditional genealogical knowledge we might be able to gather and how it might help us further. Does any of that help guide the branching in more detail? Does it help us assign ancestor names to the branching points in our genetic family tree? We have to be careful of course if the group is trying to use DNA to confirm their genealogy not to then use the genealogy information they were trying to confirm in the first place, but if their traditional research is reliable enough it may help us in our conclusions about the group's connections.

Adding data is always a good way to learn more. You can target certain upgrades, like in this example upgrading kit 90004 to at least 12 markers would help confirm whether he had the off-modal "large" for DYS011 which might help place him within the 90006, 90007, 90008, and 90009 subgroup. Or if there were higher-level tests that provided more words of the rhyme and you could get a number of these men to upgrade, that might further refine your branching knowledge.

Adding SNPs to the analysis is another obvious improvement, since SNPs can give you the basic branching structure and can quite possibly reliably break this group up into several subgroups. But for instance if you had no branching SNPs that showed you that kits 90006 through 90009 were a separate subgroup, you've at least got the STRs that help you to see that, so combining the two types of mutations is often the best way to get as much from Y-DNA analysis as possible.

You could try using tools to estimate the ages back to the common ancestors for this group – various tools can do this, including (among others) Family Tree DNA's TiP tool, the McGee utility, and SAPP. Age estimation is a complex subject but both STR and SNP-based estimation methods exist and both can only give you very general estimates for the timeframe in which the common ancestor lived, normally with an error range of at least 100 years on either side of the estimate if it's within genealogical times.

The SAPP tool also provides some automation for creating these genetic family trees even adding in SNPs and genealogy information, although it tries to build the most likely tree from the data and can't show you all the relative likelihoods of alternate scenarios, so if you do use it please consider it as a modeling tool and not a proven ancestry generator. You still have to assess the finished product yourself and decide which branches are more or less likely.

Further Resources

There are many aspects of STR analysis that I haven't covered in this simple introduction, like methods of more closely estimating the ancestral haplotype, recognizing and dealing with convergence, or adding the extra Panel 6 and 7 STRs found in Big Y700 to the analysis. But if you're just getting started with STR analysis I would suggest you leave those until you've had more practice with STRs in your project.

For more on building genetic family trees using STR mutations, the best source I can suggest is Maurice Gleeson's YouTube videos on the subject, which you should find from a simple search on YouTube using his name and "mutation history trees" (Maurice now refers to them as "genetic family trees," so you may find recent videos under that search term also). Maurice also runs the Genetic Genealogy Ireland channel on YouTube so look for some on that channel also. John Cleary also has many great videos that cover STR analysis as well so his would be another invaluable resource for further study.

Journal of Genetic Genealogy

Henry Miller, Jr. & David Alexander Miller: TMRCA and Relationship Analysis

By Mark A. Miller, Castle Rock, CO

18 DEC 2021

Purpose. Determine how the current group of YDNA test results and associated autosomal DNA data (atDNA data) support and guide further genealogical research to determine the relationship between Henry Miller, Jr. ("Henry") and David Alexander Miller ("David"), both born in the first quarter of the 19th century.

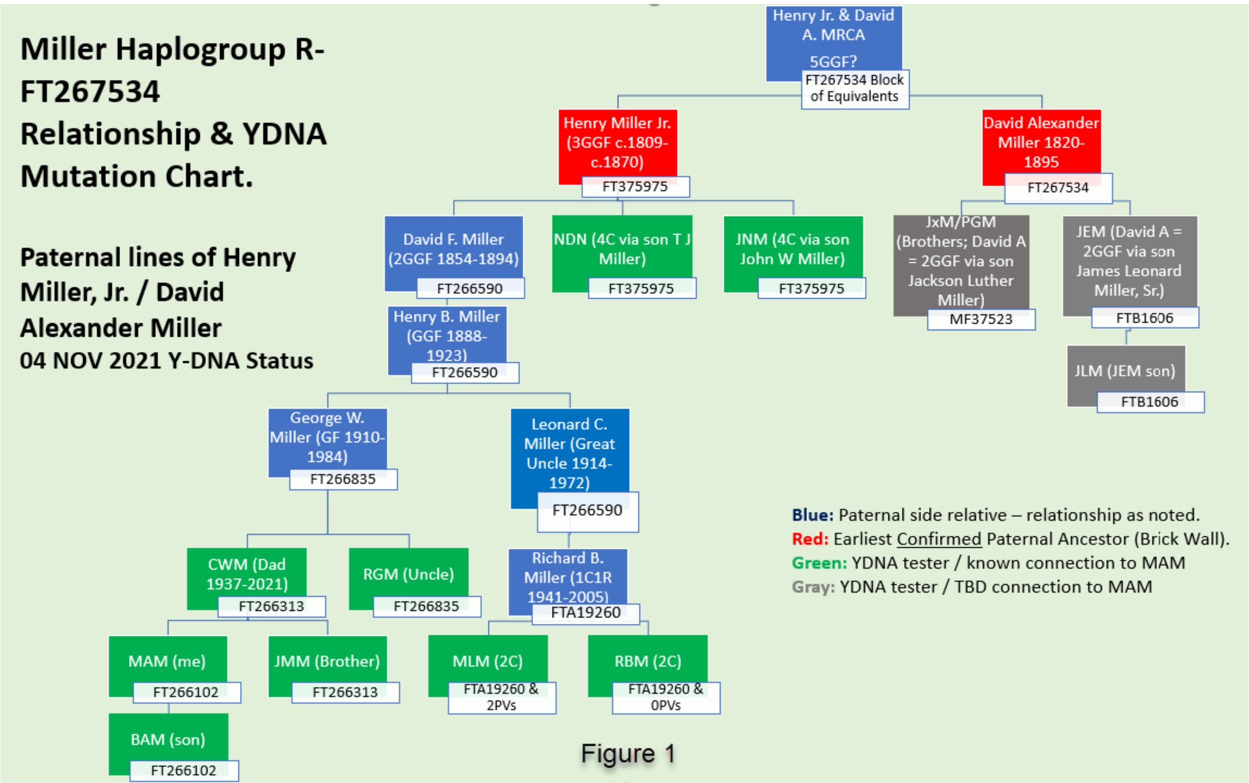

Miller Haplogroup R-FT267534
Relationship & YDNA Mutation Chart.
Paternal lines of Henry Miller, Jr. / David Alexander Miller
04 NOV 2021 Y-DNA Status
Henry Jr. & David A. MRCA 5GGF?
FT267534 Block of Equivalents
Henry Miller Jr. (3GGF c.1809-c.1870)
FT375975
David Alexander Miller 1820-1895
FT267534
David F. Miller (2GGF 1854-1894)
FT266590
NDN (4C via son T J Miller)
FT375975
JNM (4C via son John W Miller)
FT375975
JxM/PGM (Brothers; David A = 2GGF via son Jackson Luther Miller)
MF37523
JEM (David A = 2GGF via son James Leonard Miller, Sr.)
FTB1606
JLM (JEM son)
FTB1606
Henry B. Miller (GGF 1888-1923)
FT266590
George W. Miller (GF 1910-1984)
FT266835
Leonard C. Miller (Great Uncle 1914-1972)
FT266590
CWM (Dad 1937-2021)
FT266313
RGM (Uncle)
FT266835
Richard B. Miller (1C1R 1941-2005)
FTA19260
MAM (me)
FT266102
JMM (Brother)
FT266313
MLM (2C)
FTA19260 & 2PVs
RBM (2C)
FTA19260 & 0PVs
BAM (son)
FT266102
Blue: Paternal side relative – relationship as noted.
Red: Earliest Confirmed Paternal Ancestor (Brick Wall).
Green: YDNA tester / known connection to MAM
Gray: YDNA tester / TBD connection to MAM

Figure 1

Background. A team of amateur genealogists, led by Mark Miller (Colorado) and Brad Miller (Georgia), and supported by Joseph Miller (North Carolina), are utilizing traditional genealogy (atDNA testing at Ancestry, and YDNA testing at Family Tree DNA (FTDNA) to establish the patrilineal lines of descent that connect Henry and David Miller. The documented paper trail ends with each of these men, and we have been unable to establish their relationship. Henry was born in Ashe County, North Carolina in the first decade of the 19th century. David was born in neighboring Wilkes County in 1820. YDNA testing proves that Henry and David are related along their patrilineal Miller lines (Figures 1 and 4).

Because we have seemingly exhausted online genealogy resources, the team engaged local genealogist Jill Privott of Boone, NC. Ms. Privott is currently researching David's line, with a goal of identifying his connection to Henry. Based on her report, we will recruit YDNA testers to validate, or invalidate Jill's traditional genealogical research results. There are several unrelated Miller lines who migrated to this area of northwest North Carolina. The Miller surname is prevalent in the area. Identifying our Miller line from the several local Miller lines has become a process of elimination. Our strategy is to identify the most probable connection between David and Henry through traditional research, followed by YDNA testing to validate or invalidate the research. YDNA testing is effective for verifying, and for excluding, connections between patrilineal lines. In the case of a common surname like Miller, including and excluding patrilineal lines are both important elements of the investigatory process.

To date our YDNA testing has excluded the following local Miller lines as patrilineal connections to David and Henry. Each of these excluded lines is the source of many Miller descendants in the Ashe / Wilkes County area:

1) William Miller c. 1735 unknown – c. 1825, Lewis Fork, Wilkes, NC
2) Johannes Muller 1726 Ephrata, PA – 1780 Friedburg, Rowan, NC.

Objectives. Based on the 23 testers under the R-FT267000 Haplogroup (a genetically related group who share a common male ancestor), and the subgroup of 13 under R-FT267534, we estimated the following:

1) the patrilineal TMRCA (time to most recent common ancestor) of the FT267534 subclade,

2) the birth date of David's and Henry's common patrilineal ancestor, and

3) the most likely patrilineal relationship between David and Henry.

Step 1: Estimate the TMRCA of the Miller R-FT267000 Haplogroup and the average SNP mutation rate of the R-FT267534 subclade.

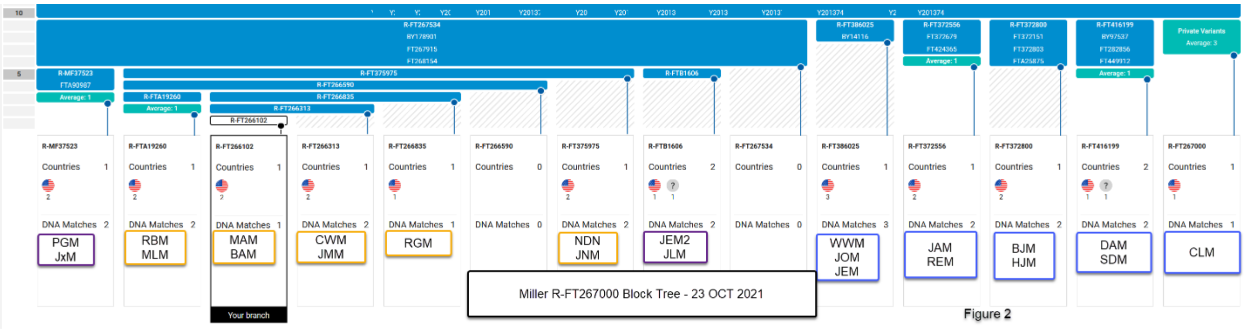


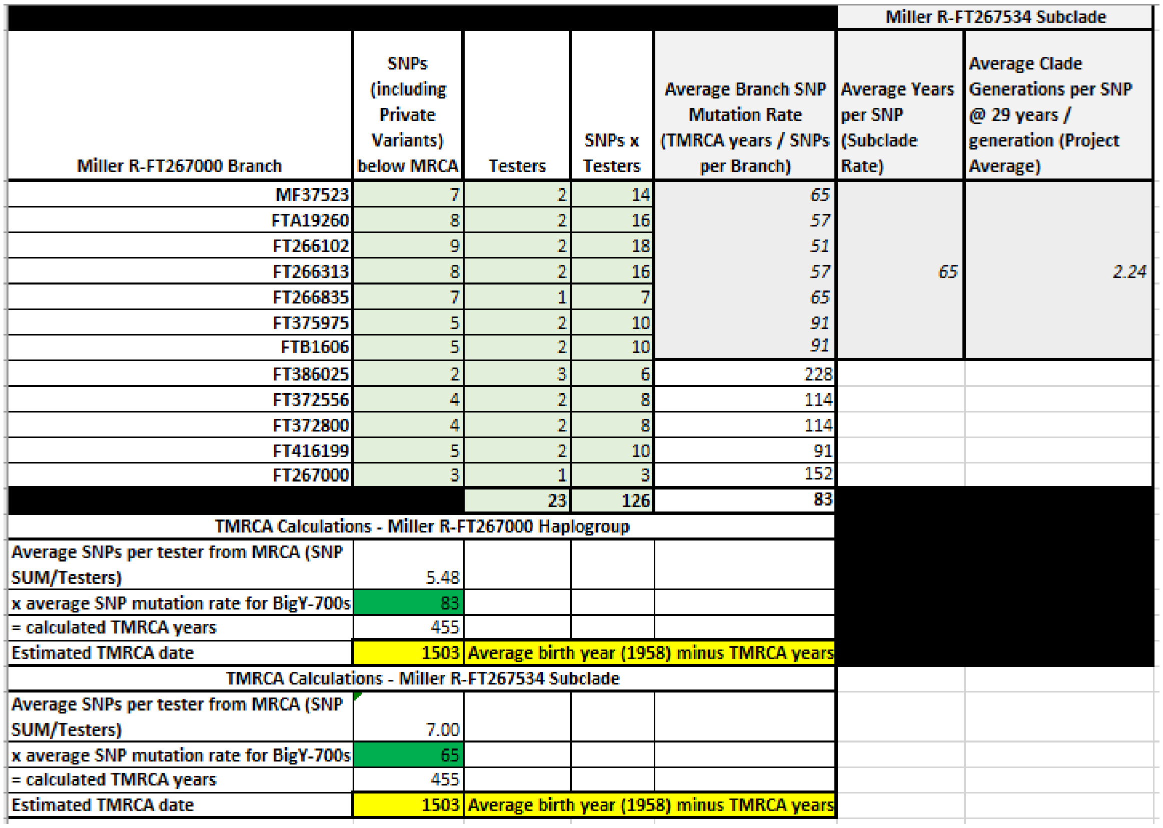
Figure 2

Figure 2 illustrates the status of the BigY-700 YDNA testing program of the Miller R-FT267000 group project through FTDNA. The average SNP mutation rate of 1 new SNP per 83 years (see https://ydna-warehouse.org/benchmarks) was utilized to calculate the project's TMRCA because all the project tests are FTDNA BigY-700 tests. There are no BigY-500 tests whose average SNP mutation rate is 131 years due to its lesser test coverage (ibid.). The author's Figure 3 spreadsheet was used to calculate the average number of SNPs per tester downstream of the project's common patrilineal ancestor. The sum of 126 SNPs divided by 23 testers equals an average of 5.48 SNPs per tester from the project MRCA to the average birth year of the FT267000 test group, which is 1958. 5.48 SNPs x 83 years/SNP equals 455 years. 1958 – 455 = 1503 AD +/-. Because TMRCA estimates are only approximate we used 1500 AD +/-.

In his review of this paper, it is noteworthy that Dr. Iain McDonald believes the R-FT267534 block may be an anomaly of 4 SNPs that occurred over a very short period, possibly even one generation, skewing the project TMRCA. To minimize the potential bias introduced by this possibility Dr. McDonald suggested:

> *"[if you] treat R-FT257534 as its own line, you end up with 43 SNPs / 13 lines = 3.31 SNPs per line. If you then average R-FT257534 and the other [Miller project] lines together (4 [in the FT267534 block]+3.31 SNP + 2+2 + 4+4 + 4+4 + 4+4 + 3 = 38.31 SNPs / 10 lines = [3.81] SNPs per line], giving an average of about [315] years [1635 AD +/-]. In reality, that could be anywhere from about [250 to 400] years, or [1550 to 1700 AD]. I'd say this is a more realistic TMRCA estimate."*

Dr. McDonald's theory suggests we may be closer to discovering the MRCA of each R-FT267000 subclade (the subgroups of the R-FT267000 Haplgroup) and the corresponding project MRCA than we believe. Pending additional testing which should indicate which approach is more likely, the author is *staying the course* with an open mind to Dr. McDonald's view, while seeking testers to break up and clarify the FT267534 block of equivalents

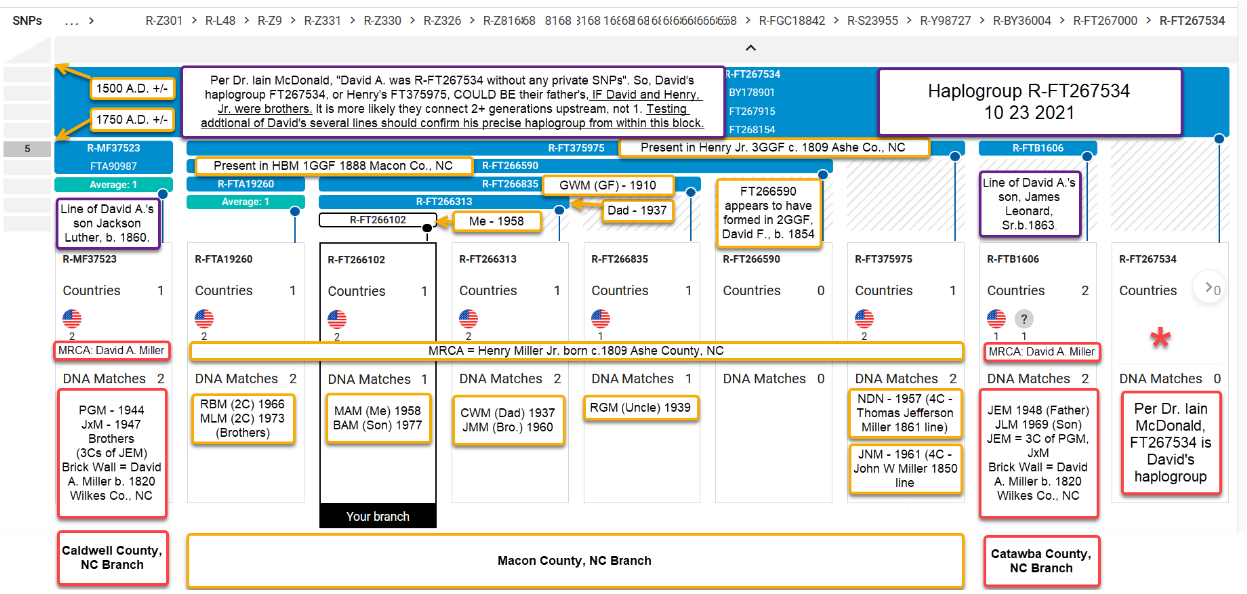


					Miller R-FT267534 Subclade	
Miller R-FT267000 Branch	SNPs (including Private Variants) below MRCA	Testers	SNPs x Testers	Average Branch SNP Mutation Rate (TMRCA years / SNPs per Branch)	Average Years per SNP (Subclade Rate)	Average Clade Generations per SNP @ 29 years / generation (Project Average)
MF37523	7	2	14	*65*	*65*	*2.24*
FTA19260	8	2	16	*57*		
FT266102	9	2	18	*51*		
FT266313	8	2	16	*57*		
FT266835	7	1	7	*65*		
FT375975	5	2	10	*91*		
FTB1606	5	2	10	*91*		
FT386025	2	3	6	228		
FT372556	4	2	8	114		
FT372800	4	2	8	114		
FT416199	5	2	10	91		
FT267000	3	1	3	152		
		23	126	83		
TMRCA Calculations - Miller R-FT267000 Haplogroup						
Average SNPs per tester from MRCA (SNP SUM/Testers)	5.48					
x average SNP mutation rate for BigY-700s	83					
= calculated TMRCA years	455					
Estimated TMRCA date	1503	Average birth year (1958) minus TMRCA years				
TMRCA Calculations - Miller R-FT267534 Subclade						
Average SNPs per tester from MRCA (SNP SUM/Testers)	7.00					
x average SNP mutation rate for BigY-700s	65					
= calculated TMRCA years	455					
Estimated TMRCA date	1503	Average birth year (1958) minus TMRCA years				

Figure 3

As shown in Figure 3, more SNP mutations have emerged down the lines of the FT267534 subclade than other associated lines. This suggests that the average SNP mutation rate of the FT267534 family clade is faster than the average 83 years/SNP factor. Over the 455-year period calculated above, the FT267534 subclade averages seven SNPs per tester or 65 years/SNP (455/7).

Step 2: Estimate the birth date of the common patrilineal ancestor of David and Henry.



1500 A.D. +/-

1750 A.D. +/-

Per Dr. Iain McDonald, "David A. was R-FT267534 without any private SNPs". So, David's haplogroup FT267534, or Henry's FT375975, COULD BE their father's, IF David and Henry, Jr. were brothers. It is more likely they connect 2+ generations upstream, not 1. Testing addtional of David's several lines should confirm his precise haplogroup from within this block.

Haplogroup R-FT267534
10 23 2021

Present in Henry Jr. 3GGF c. 1809 Ashe Co., NC

Present in HBM 1GGF 1888 Macon Co., NC

GWM (GF) - 1910

Dad - 1937

Me - 1958

FT266590 appears to have formed in 2GGF, David F., b. 1854

Line of David A.'s son Jackson Luther, b. 1860.

Line of David A.'s son, James Leonard, Sr.b.1863.

MRCA: David A. Miller

MRCA = Henry Miller Jr. born c.1809 Ashe County, NC

MRCA: David A. Miller

PGM - 1944
JxM - 1947
Brothers
(3Cs of JEM)
Brick Wall = David A. Miller b. 1820 Wilkes Co., NC

RBM (2C) 1966
MLM (2C) 1973
(Brothers)

MAM (Me) 1958
BAM (Son) 1977

CWM (Dad) 1937
JMM (Bro.) 1960

RGM (Uncle) 1939

NDN - 1957 (4C - Thomas Jefferson Miller 1861 line)

JNM - 1961 (4C - John W Miller 1850 line

JEM 1948 (Father)
JLM 1969 (Son)
JEM = 3C of PGM, JxM
Brick Wall = David A. Miller b. 1820 Wilkes Co., NC

Per Dr. Iain McDonald, FT267534 is David's haplogroup

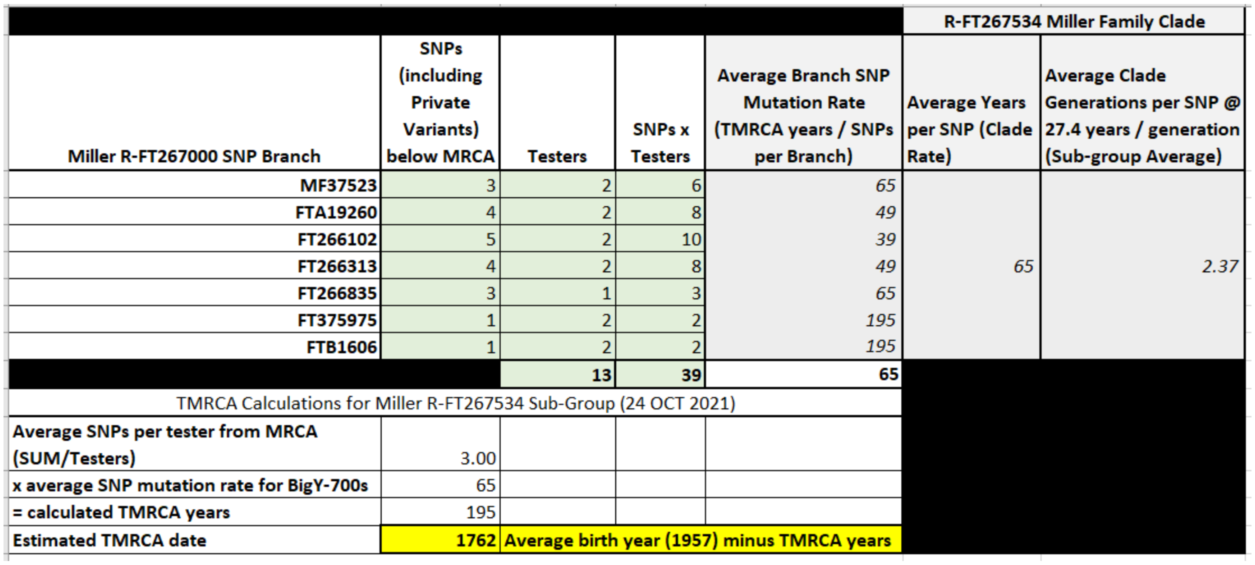
Caldwell County, NC Branch

Macon County, NC Branch

Catawba County, NC Branch

Figure 4

Again, as shown in Figure 4, there are four SNPs in the FT267534 Block of Equivalents. The gap at the base of this block represents the MRCA of the Henry and David Miller lines. To estimate the TMRCA of David and Henry these four SNPs are subtracted from the FT267534 average of seven SNPs from the project TMRCA date to present. So, the average number of SNPs per tester for the David/Henry estimate is three (7-4), as also calculated in Figure 5. Three SNPs x 65 years/SNP = 195 years. The average birth year of the testers is 1957. 1957 – 195 years = 1762 A.D. +/- as the estimated birth date of David's and Henry's MRCA.

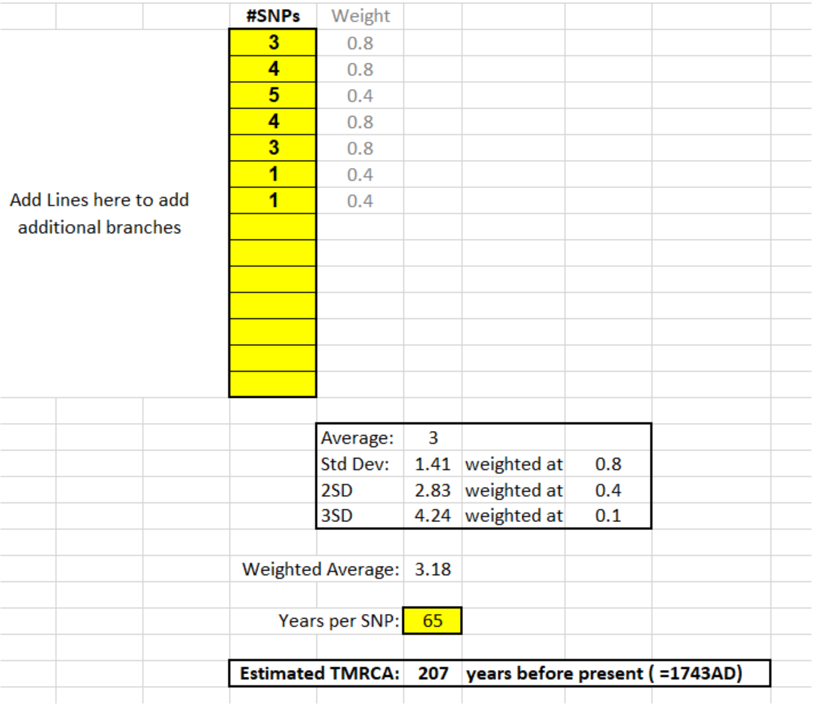


					R-FT267534 Miller Family Clade	
Miller R-FT267000 SNP Branch	SNPs (including Private Variants) below MRCA	Testers	SNPs x Testers	Average Branch SNP Mutation Rate (TMRCA years / SNPs per Branch)	Average Years per SNP (Clade Rate)	Average Clade Generations per SNP @ 27.4 years / generation (Sub-group Average)
MF37523	3	2	6	65	65	2.37
FTA19260	4	2	8	49		
FT266102	5	2	10	39		
FT266313	4	2	8	49		
FT266835	3	1	3	65		
FT375975	1	2	2	195		
FTB1606	1	2	2	195		
		13	39	65		
TMRCA Calculations for Miller R-FT267534 Sub-Group (24 OCT 2021)						
Average SNPs per tester from MRCA (SUM/Testers)	3.00					
x average SNP mutation rate for BigY-700s	65					
= calculated TMRCA years	195					
Estimated TMRCA date	1762	Average birth year (1957) minus TMRCA years				

Figure 5

David Vance's TMRCA calculator (https://docs.google.com/file/d/1Wjye3-vYDZQyNFcyR6lwK0NBmAkbc04a/edit?usp=docslist_api&filetype=msexcel) was also applied (Figure 6).

Mr. Vance’s model estimates 207 years before present (applying the 65 years/SNP mutation rate). The average birth year of 1957 – 207 years = 1750 A.D. +/- for the David A. / Henry, Jr. MRCA.



#SNPs	Weight
3	0.8
4	0.8
5	0.4
4	0.8
3	0.8
1	0.4
1	0.4

Add Lines here to add additional branches

Average:	3		
Std Dev:	1.41	weighted at	0.8
2SD	2.83	weighted at	0.4
3SD	4.24	weighted at	0.1

Weighted Average: 3.18

Years per SNP: 65

Estimated TMRCA: 207 years before present (=1743AD)

Figure 6

Step 3: Estimate the David A. / Henry Jr. TMRCA using David Vance’s SAPP application. SAPP (https://www.jdvsite.com/) is an online tool written by David Vance that builds phylogenetic trees (genetic branch diagrams) from YDNA and patrilineal genealogy data. SAPP (“Still Another Phylogeny Program”) utilizes STR data (Short Tandem Repeats - short sequences of YDNA that are repeated numerous times) and genealogical input to estimate the TMRCAs and genetic branching of a test group. SAPP does not utilize SNPs to estimate TMRCAs. SAPP applies SNP data to aid development of genetic branching. The estimating methods discussed in Steps 1 and 2 utilize SNP data; SAPP invokes STRs, a second, independent, YDNA molecular clock.

As shown in Figure 7, SAPP estimates the David/Henry MRCA convergence date to be 1750 AD, with a range of 1650 AD – 1800 AD. Assuming the FT267000 TMRCA of 1500 AD is reasonably accurate, it is unlikely that the four SNPS in the FT267534 Block of Equivalents formed between 1500 AD and the 1650 AD end of SAPP’s TMRCA estimate. That would require an average SNP mutation rate of about 37.5 years/SNP. This is possible, as Dr. McDonald points out, but it is not common. On the upper end of the range, 1800 CE is too close to the birth year of Henry, Jr. So, the error range is probably tighter on both ends.

Averaging the three MRCA estimates (1762, 1750, and 1750) equals 1754. 1755 was used as a working MRCA convergence date, with allowance for error. A 1755 MRCA estimate is consistent with Dr. McDonald’s assessment, in his review of this paper: “David was R-FT267534, without any private SNPs. No SNPs occur between David Alexander Miller and the common ancestor that unites you to him. Otherwise, JxM/PGM and JEM+JLM would have a common haplogroup that you don't have. Your common ancestor must be within one SNP timescale of 1820. That could be 83 years, or much shorter or longer. It's unlikely to be that much longer, otherwise the more distant Miller tests (WWM, JOM, JEM, ...) would show more SNPs".

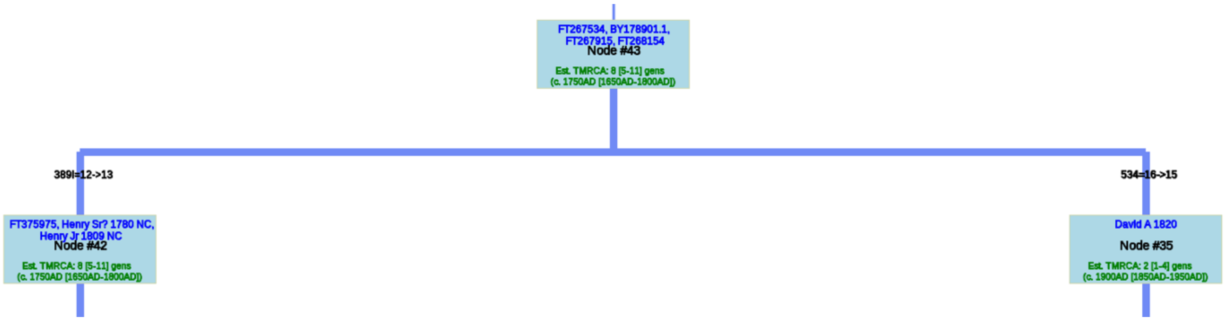
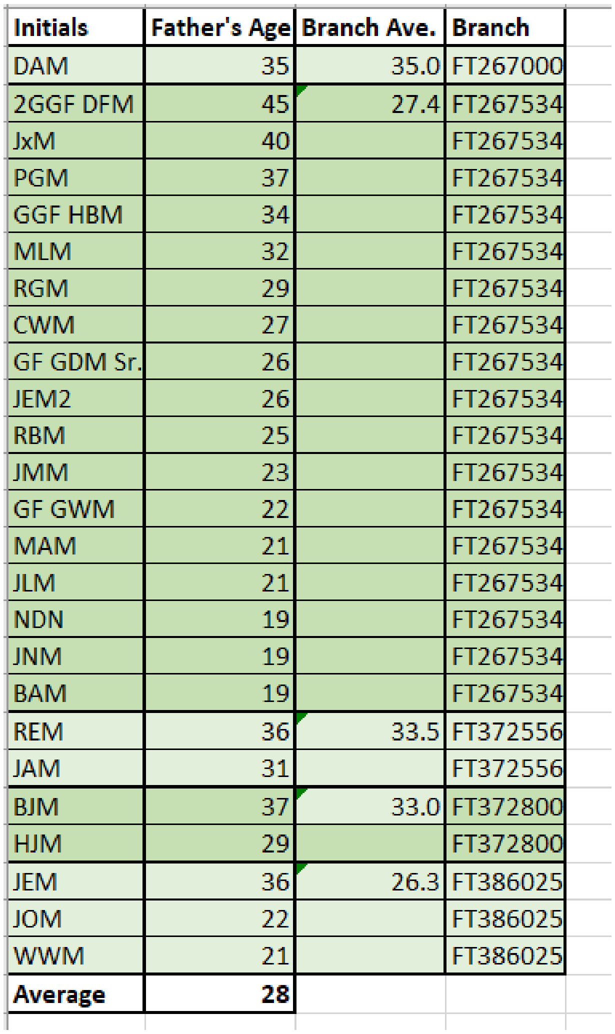

FT267534, BY178901.1, FT267915, FT268154
Node #43
Est. TMRCA: 8 [5-11] gens
(c. 1750AD [1650AD-1800AD])
389I=12->13
534=16->15
FT375975, Henry Sr? 1780 NC, Henry Jr 1809 NC
Node #42
Est. TMRCA: 8 [5-11] gens
(c. 1750AD [1650AD-1800AD])
David A 1820
Node #35
Est. TMRCA: 2 [1-4] gens
(c. 1900AD [1850AD-1950AD])

Figure 7

Step 4: Estimate the generations from David A and Henry Jr. to their MRCA.

Figure 8 calculates the average years per generation for the Miller R-FT267000 haplogroup (28 years) and the FT267534 subclade (27.4 years).



Initials	Father's Age	Branch Ave.	Branch
DAM	35	35.0	FT267000
2GGF DFM	45	27.4	FT267534
JxM	40		FT267534
PGM	37		FT267534
GGF HBM	34		FT267534
MLM	32		FT267534
RGM	29		FT267534
CWM	27		FT267534
GF GDM Sr.	26		FT267534
JEM2	26		FT267534
RBM	25		FT267534
JMM	23		FT267534
GF GWM	22		FT267534
MAM	21		FT267534
JLM	21		FT267534
NDN	19		FT267534
JNM	19		FT267534
BAM	19		FT267534
REM	36	33.5	FT372556
JAM	31		FT372556
BJM	37	33.0	FT372800
HJM	29		FT372800
JEM	36	26.3	FT386025
JOM	22		FT386025
WWM	21		FT386025
Average	**28**		

Figure 8

Applying 27.4 years per generation to estimate the number of generations from David Miller and Henry Miller to their MRCA yields:

1) David was born in 1820. 1820 – 1755 = 65 years. 65/27.4 yrs. per generation = 2.4 generations, .
2) Henry's birth year is between 1802 and 1810, say 1806. 1806 – 1755 = 51 years. 51/27.4 = 1.9 generations,

One of our working hypotheses has been that Henry and David may have been brothers. Steps 1 through 4 imply it is more likely they were 1st cousins, 1C1Rs, or 2nd cousins. If the TMRCA dates of 1500 and 1755 are valid, it seems unlikely that they would be much more distantly related than 2nd cousins, although possible.

Step 5: What can we intimate from Henry's SNP FT375975 and David's SNP FT267534?

Could David Miller and Henry Miller be brothers? Possibly. The FT267534 subclade demonstrates the ability to form SNPs at a faster than average rate. However, the 1755 MRCA implies it is more likely that

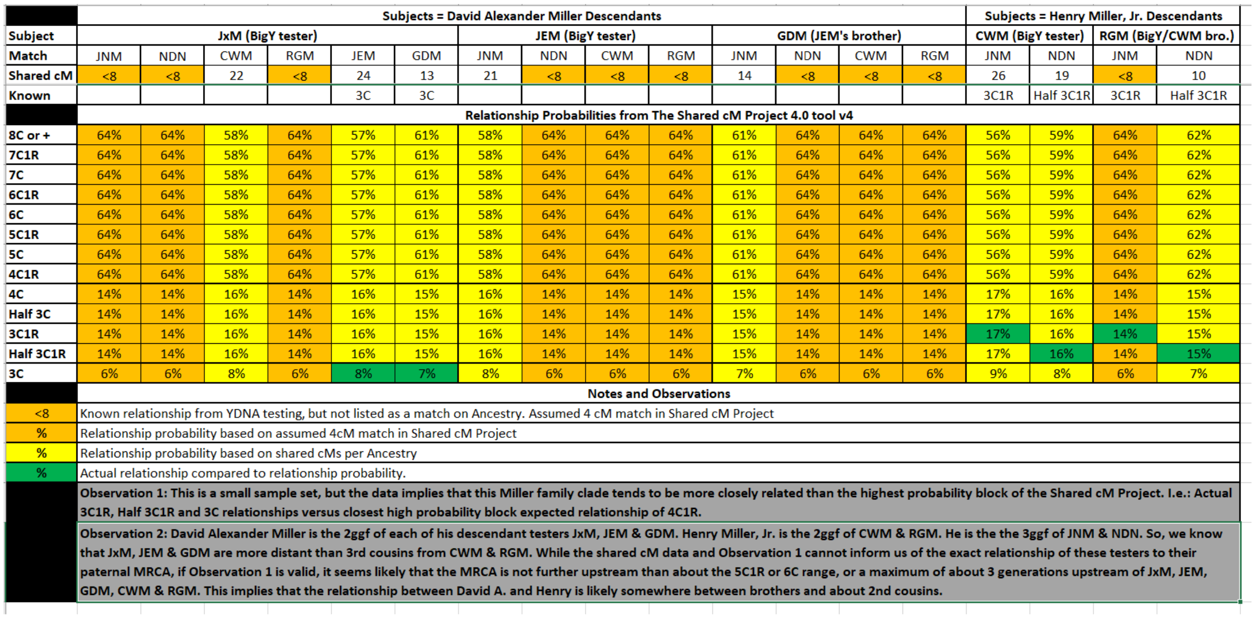
David and Henry connect two or three generations upstream, allowing for the possibility that they may have been brothers.

Step 6: What clues do atDNA relationships between the FT267534 testers (plus the brother of JEM) provide?

Figure 9 was built to identify potential atDNA patterns among the FT267534 YDNA test group. The relationship probabilities are taken from The Shared cM Project tool v4 located at dnapainter.com. As noted in the Figure, two observations emerged:

1) This is small sample set, but the data implies the FT267534 Miller family clade may tend to be more closely related than the highest probability block of the Shared cM Project, based on actual 3C1R, Half 3C1R and 3C relationships versus closest high probability block expected relationship of 4C1R.
2) David is the 2ggf of his descendant testers, JxM, JEM & GDM. Henry is the 2ggf of CWM & RGM, and the 3ggf of JNM & NDN. So, we know that JxM, JEM & GDM are at least 4th cousins from CWM & RGM. While the shared cM data and Observation 1 cannot inform us of the exact relationship of these testers to their patrilineal MRCA, if Observation 1 is valid, it seems likely that the MRCA is not further upstream than about the 5C1R or 6C range, or a maximum of about 3 generations upstream of JxM, JEM, GDM, CWM & RGM. This data also implies that the relationship between David and Henry is likely somewhere between brothers and 2nd cousins.

Does endogamy make the Miller patrilineal line relationships appear closer than actual? Henry Miller married his first wife, Mary Ann Nichols, in 1835 in Wilkes County, NC. Their first recorded child, Martha Jane, was born in 1837 in Haywood County, NC, about 135 miles to the west. So, Henry left the Ashe/Wilkes County area before the first child was born. This Miller branch never returned to the Ashe/Wilkes County area. This implies that endogamy is not a key factor between Henry's descendants and David's. Many of David's have remained in Caldwell and Catawba Counties, which are near Ashe and Wilkes Counties. While Caldwell and Catawba Counties are neighbors, examination of the family trees connecting the "Caldwell Millers" and the "Catawba Millers" does not indicate that endogamy is significant between these Miller lines. It does not appear that endogamy is significantly prejudicing Figure 9.



	Subjects = David Alexander Miller Descendants														Subjects = Henry Miller, Jr. Descendants			
Subject	JxM (BigY tester)						JEM (BigY tester)				GDM (JEM's brother)				CWM (BigY tester)		RGM (BigY/CWM bro.)	
Match	JNM	NDN	CWM	RGM	JEM	GDM	JNM	NDN	CWM	RGM	JNM	NDN	CWM	RGM	JNM	NDN	JNM	NDN
Shared cM	<8	<8	22	<8	24	13	21	<8	<8	<8	14	<8	<8	<8	26	19	<8	10
Known					3C	3C									3C1R	Half 3C1R	3C1R	Half 3C1R
	Relationship Probabilities from The Shared cM Project 4.0 tool v4																	
8C or +	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
7C1R	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
7C	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
6C1R	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
6C	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
5C1R	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
5C	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
4C1R	64%	64%	58%	64%	57%	61%	58%	64%	64%	64%	61%	64%	64%	64%	56%	59%	64%	62%
4C	14%	14%	16%	14%	16%	15%	16%	14%	14%	14%	15%	14%	14%	14%	17%	16%	14%	15%
Half 3C	14%	14%	16%	14%	16%	15%	16%	14%	14%	14%	15%	14%	14%	14%	17%	16%	14%	15%
3C1R	14%	14%	16%	14%	16%	15%	16%	14%	14%	14%	15%	14%	14%	14%	17%	16%	14%	15%
Half 3C1R	14%	14%	16%	14%	16%	15%	16%	14%	14%	14%	15%	14%	14%	14%	17%	16%	14%	15%
3C	6%	6%	8%	6%	8%	7%	8%	6%	6%	6%	7%	6%	6%	6%	9%	8%	6%	7%
	Notes and Observations																	
<8	Known relationship from YDNA testing, but not listed as a match on Ancestry. Assumed 4 cM match in Shared cM Project																	
%	Relationship probability based on assumed 4cM match in Shared cM Project																	
%	Relationship probability based on shared cMs per Ancestry																	
%	Actual relationship compared to relationship probability.																	
	Observation 1: This is a small sample set, but the data implies that this Miller family clade tends to be more closely related than the highest probability block of the Shared cM Project. I.e.: Actual 3C1R, Half 3C1R and 3C relationships versus closest high probability block expected relationship of 4C1R.																	
	Observation 2: David Alexander Miller is the 2ggf of each of his descendant testers JxM, JEM & GDM. Henry Miller, Jr. is the 2ggf of CWM & RGM. He is the the 3ggf of JNM & NDN. So, we know that JxM, JEM & GDM are more distant than 3rd cousins from CWM & RGM. While the shared cM data and Observation 1 cannot inform us of the exact relationship of these testers to their paternal MRCA, if Observation 1 is valid, it seems likely that the MRCA is not further upstream than about the 5C1R or 6C range, or a maximum of about 3 generations upstream of JxM, JEM, GDM, CWM & RGM. This implies that the relationship between David A. and Henry is likely somewhere between brothers and about 2nd cousins.																	

Figure 9

Summary of observations.

1. The TMRCA of the Miller R-FT267000 project is approximately 1500 AD, with acknowledgment of Dr. McDonald's hypothesis that the TMRCA may be in the 1550 to 1700 AD range. For determining the relationship between Henry and David the difference between these hypotheses is less important than the TMRCA for the R-FT267534 subclade and should resolve with further testing and genealogical research.
2. The TMRCA of the Miller R-FT267534 sub-clade is approximately 1755.
3. The current haplogroup status and block tree structure of the David and Henry lines implies it is not likely they were brothers, but the possibility cannot be excluded.
4. Using 27.4 years per generation for the FT267534 family group:
    a. From David's 1820 birth date to the 1755 TMRCA is about 2-3 generations.
    b. From Henry's 1806 +/- birth date to the 1755 TMRCA is about 2 generations.
    c. These approximations imply that the TMRCA is Henry's grandfather, who would also be David's grandfather or great-grandfather, implying they are cousins or 1C1rs.
        i. Assuming Henry's birth year is about 1806 and the MRCA's is 1755:
            1. If the MRCA is his grandfather, the birth year of Henry's father is roughly 1780.
            2. If the MRCA is his great grandfather, the birth years of his grandfather and father are roughly 1872 and 1889, or an average of 17 years old each when the grandfather and father were born.
            3. The first scenario is more likely than the second.
        ii. Applying David's birth year of 1820 and the 1755 TMRCA:
            1. If the MRCA is his grandfather, the estimated birth year of his father is roughly 1788

        2. If the MRCA is his great-grandfather the birth years of his grandfather and father are roughly 1777 and 1798, or an average of 22 and 21 years old when the grandfather and father were born.
        3. Both possibilities are reasonable.
5. atDNA analysis implies that the likely relationship between David and Henry is in the range of brothers to 2nd cousins.

Conclusion. Recognizing that there is a margin of error, the most likely relationship between Henry and David is 1C1Rs or 1st cousins, with a 2nd cousin relationship possible. Pending further testing and research that supports a different theory, the hypothesis that 1C1R plus or minus is the most probable relationship between David and Henry because it most effectively reconciles the body of data:

1) The proximity of their birth locations and dates.
2) The migration history of northwest North Carolina, as briefly discussed below.
3) The estimated 1755 TMRCA compared to the generational analysis.
4) The block tree and atDNA data suggest they are closely related, but not brothers.

Migration south along the Great Wagon Road began prior to 1750. It appears the Miller family migrated to northwest North Carolina sometime after the conclusion of the French and Indian War in 1763, and prior to Henry's birth in the first decade of the 1800's. It is likely they migrated in the late 1700's as Henry's Miller line did not become well established in the area, and David's only became well established through his 10 children (seven sons and three daughters) who largely remained in the area. Whether the original patrilineal ancestor(s) migrated to northwest North Carolina with his family, or his descendants migrated on their own, the conclusions of this paper reconcile well with the birth dates and locations of Henry and David. The data support the hypotheses that the family migrated to Wilkes County prior to the 1799 creation of Ashe County. There is evidence that one of the branches of the family either moved, creating the Ashe and Wilkes birth locations, or the creation of Ashe County placed one branch in Wilkes and the other in newly formed Ashe County. Either option is possible.

Next steps.

1. After receipt of the Privott genealogy report, update this analysis, if needed.
    a. Define the next YDNA testing plan.
    b. Add one or more YDNA testers with a potential upstream connection to David and Henry, to test the report.
2. Add at least one YDNA tester down one or more of the untested lines of David's descendants to empirically define David's haplogroup.
3. Find one or more testers to validate or exclude the Eli Jackson Miller line as a third connection to the David and Henry lines.
4. Apply Dr. McDonald's advice from his review of this paper:
    a. "The other important thing is to consider the more distant Millers in the tree. The overall haplogroup can't be that old, otherwise people would have different surnames. This acts as a safe upper limit for the age of the overall R-FT267000 block", and
    b. "I would also pay attention to these more distant lines and see whether you can tie together a more distant line to any of these. Get any one of them back to the early 1700s or 1600s, and you may find your own ancestors among them."

I would like to express my sincere gratitude to Dr. Iain McDonald, Dr. Lee Martinez, and David Vance for reviewing this paper and providing insightful feedback that forced deeper thought. Not all review comments were incorporated, but all are greatly appreciated and may prove to be more correct than my conclusions as genealogical research and further testing bring the Miller team closer to the truth. The author is solely responsible for the accuracy and completeness of this paper.

Journal of Genetic Genealogy

AUTOSOMAL DNA AND GENEALOGY
How Can DNA and Triangulation Be Used to Identify, Evaluate and Present Conclusions of Relatedness?

by

Thad Thomas, MSc

June 2016

Acknowledgements:

Producing a dissertation is a monumental production. This one has been no exception. I have been the beneficiary of many miracles—great and small—in my journey to discover, synthesize, and present this subject matter.

I give thanks to my cousins who have invested in and shared their genetic heritage, enabling my journey in the wilderness of genetic genealogy. I acknowledge the expertise of citizen scientists and academics who have made their findings accessible—oases enabling the journey. I thank family, my supervisor, and friends who have given of their time to read and comment on my work as I have sought the paths others have followed and attempted to blazed my own trails in this wilderness.

My wife and children have shouldered the biggest burdens during this journey. I humbly and gratefully acknowledge their many sacrifices.

ABSTRACT

DNA genotyping allows consumers to examine their genetic heritage, including (within limits) a record of their ancestry. Computer algorithms can associate genotyped individuals who appear to share a genetic past, but appearances cannot be the basis for declaring genealogical relationships. While parent/child relationships can be established with certainty, other genealogical relationships can only be estimated—even in cases where a genetic relationship is sure to exist. How, then, can autosomal DNA (atDNA) be used to reveal, confirm, or even prove genealogical relationships?

This text presents a methodology for identifying, evaluating, and presenting a conclusion of relatedness utilizing atDNA. This methodology fits firmly within the framework of the genealogical research process, enabling atDNA to be used in proof arguments about relationships through the processes of question asking, information gathering, hypothesis testing, conclusion accepting, and proof explained.

The genetic genealogical community has not been definitive about acceptance criteria for atDNA-based conclusions (i.e., required elements, standards of evaluation, etc.), and more especially triangulated conclusions. It is the author's opinion that triangulation ought to be central to most genetic genealogical proof arguments, yet the literature explaining triangulation fails to make plain the reasons why triangulation is effective. These deficiencies can be addressed by decomposing triangulation into its fundamental building blocks and re-presenting it in the context of the genealogical research process.

This text focuses particularly on how hypothesis testing is used to determine strengths and weaknesses of atDNA-based conclusions. This discussion includes important heuristics that help the genetic genealogist understand the capabilities and limits that accompany this biotechnical genealogical record.

CONTENTS

LIST OF FIGURES

DEFINITIONS

Allele In the context of a SNP, an allele is one of two or more alternative values that may be found when assaying a SNP.[1]

Assay A procedure for qualitatively assessing or quantitatively measuring the presence or amount of a target entity.[2]

atDNA Autosomal DNA. The autosomes are all of the chromosomes (numbered 1 through 22 from longest to shortest) that are not the sex chromosomes (the X and Y chromosomes).

Base pair A pair of nucleotides (bases). DNA molecules incorporate four nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T). In DNA, these nucleotides always form up in pairs—A always pairs with T, G always pairs with C—and thus the designation "base pair".

Chr1, …, Chr22 A shorthand for referring to the numbered autosomes—i.e., Chr1 is Chromosome 1, and so on.

cM Centimorgan. A unit of genetic distance. 1 cM corresponds to a 1% chance of a recombination (crossover) event occurring between two locations on the chromosome.[3]

Diploid Regarding cells, refers to the fact that each contains two complete copies of the genome—one copy from each biological parent.[4] See also: *haploid*.

DNA Deoxyribonucleic acid.

Genome A complete haploid set (a single copy) of the chromosomes (genetic information) held by a gamete, microorganism, or multi-cellular organism.[5,6] Humans are diploid and, as such, have two haploid copies of the human genome in each somatic cell—one copy from each biological parent.

Genotype A set of genetic markers (e.g., SNPs or STRs) selected for their likelihood of variation and used to study the genetic makeup of individuals—usually used as a proxy for the whole genome of an individual.[7,8]

HIR Half-identical region. In a diploid subject, a region along a homologous chromosome pair where only one of the two alleles from that pair matches only one of the two alleles from another subject's pair across an entire region (segment), though allowances for differences due to a mutation are generally included. See also: *match*, *segment*.

Haplotype The genotype data for a single chromosome. See also: *genotype*.

Haploid Regarding cells, refers to having a single set of unpaired chromosomes.[9] See also: *genotype*.

IBC Identical by chance. A characterization applied to a *match* when at least one of the sequences used in the comparison does not actually correspond to a real sequence found on either haploid chromosome of the haplotype from which it originated—usually the result of a computer algorithm processing unphased haplotype data. Though this concept seems useful, it is not used extensively in the literature; it is generally lumped into the more generic IBS concept. See also: *IBS*.

IBD Identical by descent. Also: identity by descent. A characterization applied to a *match* when the sequence shared between two or more haplotypes is identical because it was inherited from a recent common ancestor.

IBS Identical by state. A characterization applied to a *match* when the haplotype sequence is identical for reasons other than inheritance by descent; not IBD. Some reasons a *match* may be IBS are because it is IBC, or because it is ancient DNA (common to a population and no longer distinguishable as IBD from a recent ancestor).

ICW group In Common With (ICW) group. Given two persons who match each other, and the list of matches associated with each person, it is the set of matches that is common to both match lists.

Locus (loci) A specific location or position.[10] Each SNP is situated at a known locus. The plural form of the word is *loci*.

Mash-up A mixture or fusion of disparate elements; something created by combining elements from two or more sources.[11,12]

Match Also called: matching segment; shared segment; identical segment. When comparing two haplotypes, an assertion that at least one of the two alleles from one haplotype matches at least one of the two alleles from the other haplotype across an entire region (segment), though such assertions generally include an allowance for mutations. See also: *HIR*, *segment*.

Mbp Megabase pairs or million base pairs. A unit of physical length along a DNA strand.

Meiosis A type of cell division that produces reproductive cells (egg or sperm) in which the diploid chromosome is reduced to a haploid.[13],[14]

MRCA Most recent common ancestor.

NPE Several different expansions of this acronym are present in the literature. Here are a few:[15]

- non-paternity event
- non-paternal event
- non-parental event
- not the parent expected

In all cases, it refers to cases of misattributed paternity.

Nucleotide A molecular building block (monomeric component) of the polymers DNA or RNA.[16] Also called a “base”.

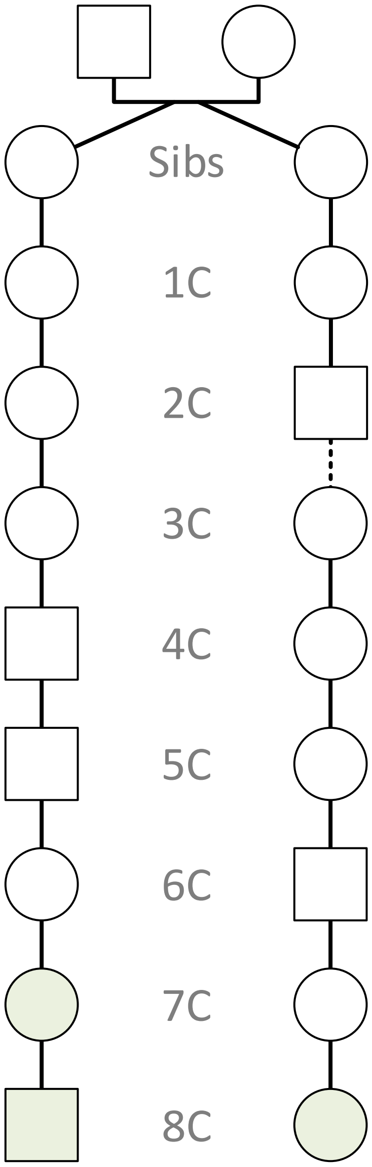
Segment A region of a homologous chromosome pair for diploid cells, or a region of a single chromosome in the case of a haploid cell. See also: *match*, *HIR*.

SNP Single nucleotide polymorphism. A single nucleotide variation that occurs at a defined location within the genome.

STR Short tandem repeat.

Triangulation A technique that involves finding three (and preferably more) persons that have a portion(s) of their DNA that is identical in a way that suggests they share a common ancestor, and then correlating this suggested commonality with genealogical research that shows a common ancestor is uniquely shared among these same persons.

KEY FOR LINEAGE DIAGRAMS



Sibs	Siblings [Often spelled out, but abbreviated in a few cases.]
1C - 8C	1st cousin, 2nd cousin, …, 8th cousin, …
└─┘	couple relationship
| or ⋮	parent/child relationship (dashed is proposed)
□	male relation
○	female relation
🟩	shading indicates genotyped individuals

INTRODUCTION

DNA genotyping gives consumers an opportunity to discover their genetic heritage, giving them a personal profile—a genotype—made up of hundreds of thousands of genetic markers.[17] In genetic genealogy, genotypes are used to associate individuals that may share a genetic relationship. Yet many of the suggested relationships are mirages—a by-product of data and algorithm limitations. Even in cases where a genetic relationship is sure to exist, the exact nature of that relationship can only be estimated. For example, one genotyping company estimates the relationship with a grandmother and her grandson as a range that includes siblings, aunts/uncles, nieces/nephews, grandparents/grandchildren, half-siblings, first cousins, great aunts/uncles, great grandparents/grandchildren and half aunts/uncles.[18] To what extent, then, can autosomal DNA (atDNA) genotypes be used to reveal and confirm (even prove) genealogical relationships?

This text investigates core concepts and methods that enable genetic genealogists to make use of atDNA to reveal and/or corroborate genealogical relationships—adding another type of genealogical record to the repertoire that researchers can use to solve genealogical problems.

This text is addressed to experts in the genetic genealogical community, and more particularly to those experts who promote the use of atDNA in genealogical research and who teach genealogical enthusiasts how to make use of their genotypes to address genealogical research questions. It is the author's hope that these experts will seek standard ways to identify, evaluate and present genetic genealogical conclusions, and that the methodological framework presented herein can be instrumental in bringing the community toward a consensus in these matters. It is the author's belief that genealogists with a serious interest in using atDNA to further their genealogical research will also benefit from this text.

GENOTYPING

Genotyping atDNA has been available direct-to-consumers since 2008.[19] In 2016, three genotyping companies have large consumer genotype databases: AncestryDNA™, 23andMe™ and Family Tree DNA™. Most of their genotypes are from US-based consumers.

Genotypes from AncestryDNA™, 23andMe™ and Family Tree DNA™ are proprietary compilations of the assayed values of single nucleotide polymorphisms (SNPs). Genotypes from these companies are based on the examination of 570,000-700,000 SNPs.[20]

So-called "next-generation" sequencing services are moving toward broad availability, making complete genome sequencing a reality.[21] Genetic genealogists do not have broad access to such data yet, so this text does not focus on the potential usefulness of such data.

Milestones in the Development of atDNA Genotyping	
1871	Friedrich Miescher published a paper that identified the presence of DNA in the cell nucleus.[22]
1904	Walter Sutton and Theodor Boveri had independently proposed that chromosomes were involved in inheritance and behaved in accordance with Mendelian laws, but it took until 1915 for their theories to be fully accepted.[23]
1953	James Watson and Francis Crick proposed the double helix structure of DNA along with the idea that it could be separated into strands to be copied.[24]
1977	Walter Gilbert and Frederick Sanger independently developed rapid DNA sequencing techniques that could be used to sequence genes.[25]
1983	Kary Mullis developed the polymerase chain reaction (PCR) to amplify DNA.[26]
1985	Alec Jeffreys developed a mechanism for profiling a genome.[27]
1990	The Human Genome Project was launched to sequence the entire human genome.[28]
2000 (March)	Family Tree DNA™ began offering a service to produce a 12-marker Y-chromosome haplotype.[29]
2000 (June)	An announcement was made concerning the initial completion of a full sequence of the human genome.[30]
2008	Genotyping companies (including 23andMe™) began selling direct-to-consumer (DTC) atDNA genotyping services.[31]
2010	Family Tree DNA™ began a phased roll-out of their atDNA genotyping service.[32]
2012	AncestryDNA™ launched their atDNA genotyping service.[33]

Figure 1: Milestones in the development of atDNA genotyping.

Genetics and Genomics

Genetic genealogy has its foundations in genetics—a specialty within molecular biology that is concerned with heredity, its mechanisms, and variation of inherited characteristics.[34] Geneticists tend to focus on genes with known functions.[35] As computing and genome sequencing have advanced, a new specialty—genomics—has emerged: the study of the totality of an organism's genetic material.[36]

Understanding the fundamentals of genetic inheritance is critical to making use of genomic information for genealogical purposes. Figure 2 depicts the pattern of inheritance for the recombining (autosomal pairs) and non-recombining (mtDNA and Y chromosome) parts of the genome.

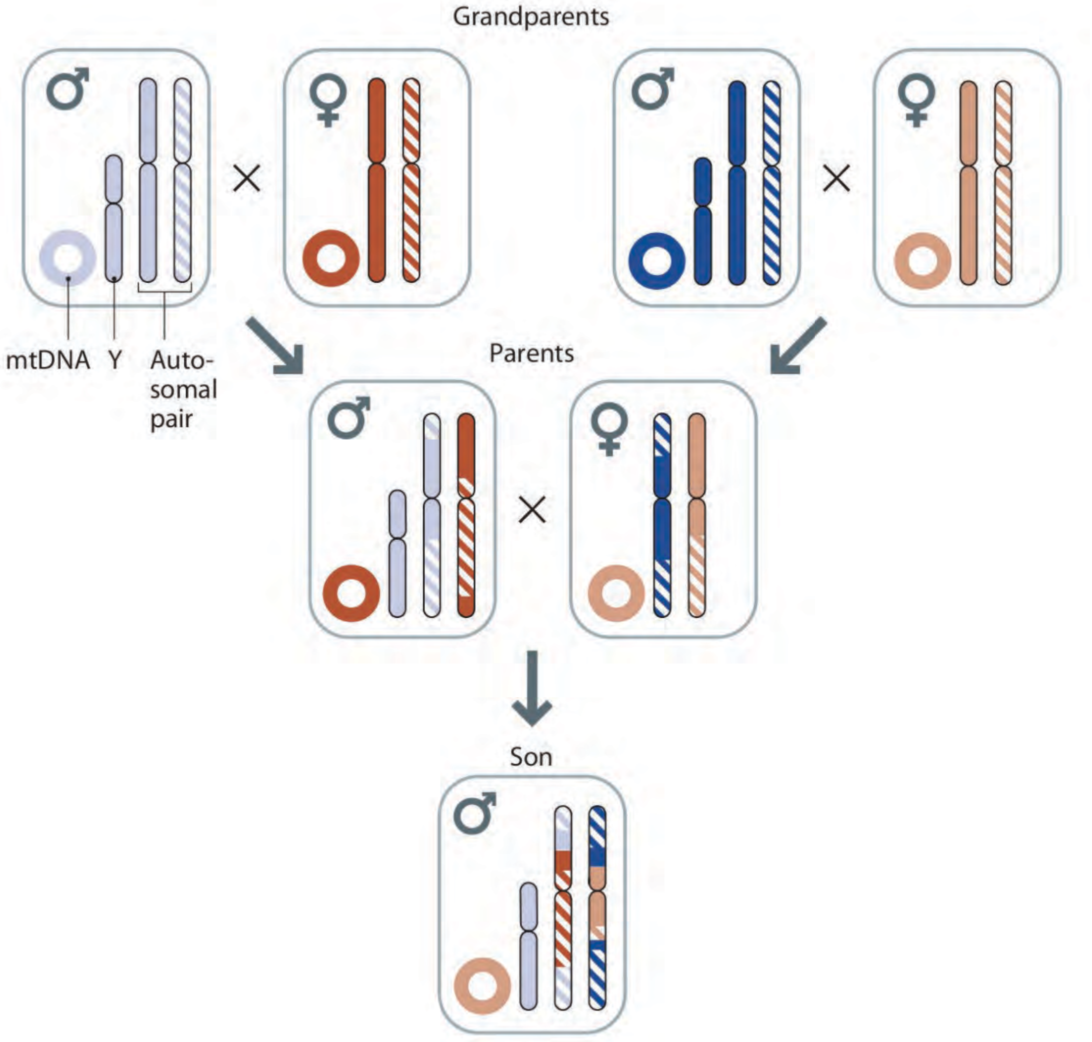


Figure 2: Inheritance of recombining and non-recombining portions of the genome.[37]

Mitochondrial DNA (mtDNA) is passed without recombination from a mother to her children (male or female). When considering an individual, it is known that they received their mtDNA from their mother, who received it from her mother, and so on, back to Mitochondrial Eve.

©2021 This work is licensed under a CC BY-NC-SA 4.0 license.

Most of the Y chromosome (or yDNA)—about 90% of it—is passed without recombination from a father to his son(s).[38] When considering an individual male, he received his yDNA from his father, who received it from his father, and so on, back to Y-Adam.

The autosomes—a focus of this text—are all the chromosomes (numbered 1 through 22 from longest to shortest) that are not the sex chromosomes (the X and Y chromosomes). Autosomes are paired—diploid. Given an individual, one chromosome in each pair was received from their father (the haploid paternal chromosome), and one from their mother (the haploid maternal chromosome). During meiosis, material from both chromosomes in a diploid pair can be reshuffled (a process called recombination) to form a unique, new haploid copy of the chromosome—a mash-up of material from the paternal and maternal haploid chromosomes. The result of this reshuffling has been represented in Figure 2.

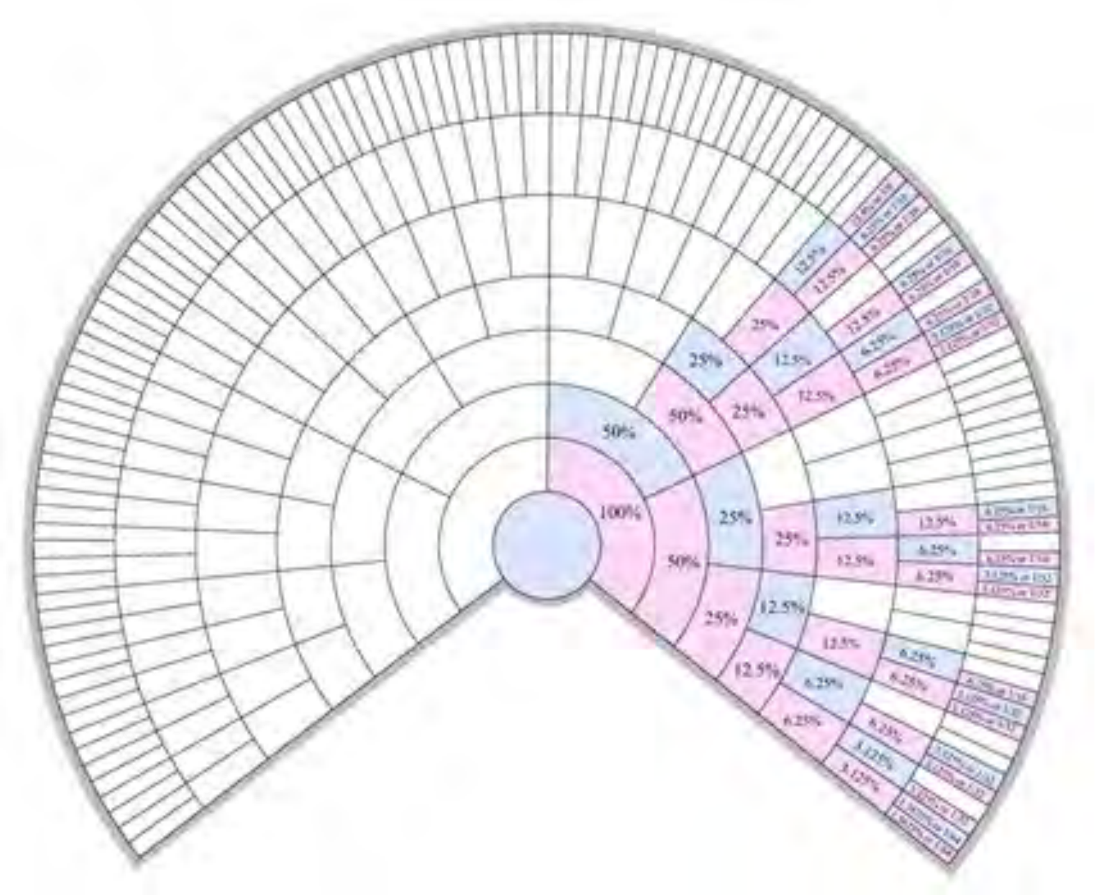


Figure 3: Male inheritance of X DNA.[39]

The X chromosome is not specifically depicted in the Figure 2 diagram (and many diagrams like it). It is another portion of the genome that undergoes recombination.

While it is possible that some recombination occurs when an X and Y are paired (the sex chromosome configuration in males), it is the reshuffling that occurs when two X chromosomes are paired (the sex chromosome configuration in females) that is generally responsible for variation in a haploid X chromosome.[40] Thus, the paternal copy of the X chromosome received by a daughter is received essentially unmodified from her father. On the other hand, the maternal copy of the X chromosome received by a child (son or daughter) from their mother is a reshuffled copy of their mother's X DNA. *Appendix E* examines an X chromosome match.

The primary source of variation in the autosomes (and X chromosomes) is recombination. Recombination rates are higher in women than men—researchers reporting about 41 crossovers per meiosis in women compared to about 27 crossovers in men.[41,42] There are recombination hotspots—loci on chromosomes where crossover events are more likely. Other types of variation in the autosomes are rare; the mechanisms that replicate and maintain DNA are very resistant to mistakes.

LITERATURE

This text makes heavy uses of material from blogs and online forums. Proving genealogical relationships using autosomal DNA in general, and triangulation in particular, has not received attention in peer-reviewed journals. Much of the discussion around these topics plays out online in these mediums, among citizen scientists and community pundits alike.

Personal Genomics

Also called recreational genomics, personal genomics is an area of emerging consumer interest. People are curious about what their DNA says about their health, physical characteristics, and ethnicity.[43] Media and advertising around DNA genotyping (particularly atDNA genotyping) caters directly to these common interests.[44,45]

DNA genotyping promises to help clinicians optimize health care, to answer questions about ancestral origins and give insights into personal physiology.[46,47] However, interpretation of genomic information is in its infancy. Science and its ability to prognosticate are always ahead of its ability to intervene or affect outcomes.[48]

As the cost of sequencing individual genomes falls, studies that incorporate full genome sequencing are on the rise.[49] With consumer confidence reasonably high, research groups are seeking to engage consumers interested in personal genomics, asking them to share their genomes to accelerate genetic research objectives.[50,51]

Social issues can affect consumer confidence. Of particular concern is information protection—keeping personal genomic details from being sold, misused, and out of unfriendly hands.[52] One company suspended access to personal genomic data when a firestorm of media attention became a public relations nightmare.[53,54]

DNA Inheritance

Solving genealogical problems using atDNA has its foundation in genetics. Genealogists desiring to make use of atDNA require fundamental knowledge of DNA inheritance—yDNA, mtDNA, and atDNA (including the X-chromosome inheritance)—and a sense of the types and sources of variation in the genome (particularly in the autosomes). This knowledge is central to reasoning about DNA and its relevance to genealogical problems. The volume of literature exploring and explaining these topics is enormous and not a focus of this text.

The International Society of Genetic Genealogy (ISOGG) hosts a wiki encyclopedia detailing many of the topics important to DNA inheritance.[55] This wiki is a recommended resource in the genetic genealogy community. Although ISOGG requires wiki contributors to make application for wiki edit privileges, content is susceptible to the same foibles common to all crowd-sourced content: inaccuracies, lack of depth/detail, missing attribution, and/or lack of curation by subject matter experts.[56,57] Information collected there is valuable but should be interpreted with appropriate caution. Researchers should consult the work of credentialed experts (e.g., Jobling, et al, or Dudley and Karczewski) to fully understand these topics.[58,59] The deeper understanding gained by forays into the academic literature will help them as they make decisions and form conclusions using genetic principles.

SUCCESS STORIES

In a study published in 1998, DNA was used to show that US President Thomas Jefferson fathered at least one of the children of his slave Sally Hemings.[60] In 2009, DNA genotyping (including atDNA genotyping) helped confirm the murder of two missing children of Russian Tsar Nicholas II.[61] Archaeologists, excavating an abbey found under a car park in 2012, found a skeleton that was later shown to be Richard III, King of England.[62,63] These peer-reviewed findings demonstrate DNA's power to prove relationships. Others, such as Richard Hill, share their personal journey to finding their biological roots.[64,65] ISOGG has created a space for members to post such success stories.[66] Plenty of literature gives efficacy to the use of DNA in establishing genealogical relationships.

PROOF

As relationships are identified using atDNA, the question that naturally follows is: Does atDNA prove the relationship?

Questions of proof using DNA dwell within the broader topic of genealogical proof. The Board for Certification of Genealogists (BCG) has been, perhaps, the most vigorous proponent for implementing genealogical standards. Its first attempt to codify a Genealogical Proof Standard (GPS) was *The BCG Genealogical Standards Manual* published in 2000.[67,68] Others, particularly Mills, have done much to bring this GPS out of obscurity and into the collective consciousness of genealogical practitioners.[69,70,71] Jones wrote a clarifying volume focused on the core elements of the GPS.[72] Further refinements have since been published in the *Genealogy Standards* manual.[73] The GPS defines foundational concepts (i.e., sources, information, evidence, hypotheses, conclusions, and proof) and describes fundamental processes and conditions—standards—by which genealogists achieve valid and acceptable outcomes. It will be shown that these concepts can be applied to atDNA in making genealogical conclusions.

When considering genealogical relationships determined by DNA analysis, the genetic genealogy community continues to debate what is required to demonstrate proof. Discussion can be found in a variety of settings (e.g., blog posts, forum discussions, conferences, mailing lists).[74,75,76,77] Some seem to want to impose so many conditions as to make conclusions an impossibility.[78] Others are quick to point out how tenuous any conclusion can be.[79] For example, in a contribution to the community's on-going discussion of triangulation, Jim Bartlett posted an article on his blog about intermediate common ancestors having a role in triangulation that sparked a number of discussions.[80] Jason Lee posted a three-part article on triangulation in response to Bartlett's post, but his discussion of triangulation does not include the concepts presented by Bartlett—giving Bartlett's ideas no credibility.[81] In a post to the ISOGG Facebook group, Blaine Bettinger lists several triangulation reliability factors important in his thinking and adds Bartlett's concept as possibly having merit.[82] Kathy Johnson seems much more adamant about the need for intermediate common ancestors, discounting triangulation solutions that do not include them.[83] Principles that can guide genetic genealogists in forming and qualifying their conclusions are developed in the *Methodology* portion of this text.

Social Concerns

Genotyping brings with it a host of social issues. Scientists are now able to modify DNA to achieve a desired outcome (i.e., genetic engineering).[84] Even if only to prevent a known genetic disorder, such capabilities raise many ethical questions.[85] The notion that genetics are the primary contributor to personal characteristics, including intelligence, behavior and health—sometimes called genetic determinism—led high-profile actress Angelina Jolie to choose elective surgery in hopes of avoiding cancer.[86] Movies such as *Gattaca* highlight fears that misuse of DNA information will lead to discrimination and other social problems.[87] Trepidation about privacy and custodianship are among a host of other concerns.[88] The Genetic Genealogy Standards Committee is proactively creating standards that include the need to address social and ethical concerns.[89]

Alongside these broader concerns are anxieties and consequences affecting individuals and/or families. DNA can reveal secrets that individuals involved may wish had remained hidden. Birth parents may wish to remain anonymous or may be unwilling to admit their role in a child's birth. These unexpected results can cause turmoil and pain for those involved.[90] One should consider potential social pitfalls, and even implement pre-emptive measures, as they undertake genetic genealogical research.

QUANTITATIVE ANALYSIS

Analyzing the total amount of atDNA shared between two individuals compared to an "expected" amount—something Tim Janzen calls quantitative analysis—is important both as a means of estimating relationships and as a means of supporting conclusions.[91] Tables and diagrams expressing the mathematical halving of the expected amount of atDNA received by each succeeding generation abound.[92] This technique is effective in identifying close family relationships—to fourth degree relationships on Bettinger's scale (see Figure 10).[93,94]

There is a clear disconnect between the small "expected" amounts and "observed" amounts as relationships become more distant.[95] Geneticists have identified this as an expected result.[96,97,98,99,100] Yet, many genetic genealogists struggle to understand this disconnect. Blaine Bettinger's *Shared cM Project* gives important insight to this variance but has only published data for relationships closer than fourth cousins.[101] Speed and Balding's findings give important insight into this variance, particularly as it pertains to more distant relationships; however, a genetic genealogist will wish their work had been presented in centimorgans, and with more granularity, when considering distant relationships.[102]

Luke Jostins gives an oft-quoted answer to a related question: “What percentage, on average, of an individual’s genealogical tree at X generations is part of their genetic tree?”—estimating that the probability that one will share atDNA with all of their ancestors in a given generation starts dropping in the fifth generation.[103] Jostins now explicitly defers to Graham Coop’s work answering this same question—that the probability starts dropping in the seventh generation.[104] Jostins’ fifth-generation answer is still prominent in genetic genealogical literature.

TRIANGULATION

This text focuses particularly on a technique referred to herein as *triangulation*. Triangulation is accepted by many as a method for proving relationships and is especially relevant to establishing distant relationships.[105] At its core, triangulation involves finding three (or more) persons that share an identical atDNA segment (HIR) and that also have genealogies that show a single common ancestor is uniquely shared among them.[106,107] As most of a person’s atDNA matches will be distant cousins, triangulation is an essential tool in the genetic genealogist’s repertoire.

In the body of literature produced by the genetic genealogical community, *triangulation* is used in a few different ways, making it necessary to disambiguate this term as used in this text from other uses common in the community. For some, triangulation is a set of mechanical genotype comparisons (no genealogy required) relative to a specific matching segment used to confirm membership in a triangulated group (see *Identifying Triangulation Building Blocks* and its subsections starting on p. 55). These rote comparisons—sometimes disambiguated as *segment triangulation*—are a necessary part of the technique that is the focus of this text but are not *triangulation* as used in this text.[108]

Sometimes genealogists seem to equate *In Common With (ICW)* groups (see *Definitions*) with triangulated groups—associating *triangulation* with attempting to identify relationships using ICW comparisons.[109] Researchers, particularly in the adoption community, have used ICW groups to great effect. One of the reasons for this success is that many of these groups would, in fact, triangulate if the relevant details of these groups could be known. ICW analysis can function as a proxy (of sorts) for triangulation, a tool to use when the information required for triangulation is not available. However, ICW comparisons are not sufficient to assure triangulation. Because there is more than one possible genetic relationship within an ICW group, what can be concluded by using these groups seems ill-defined.

Genotyping companies provide varying levels of support for triangulation, including providing no support.[110] Deficiencies inherent in genotyping company tools can be overcome using third-party tools (e.g., GEDmatch.com) if one's matches will cooperate.[111]

Triangulation begins with identifying a triangulated group (TG) and culminates with a conclusion about a common ancestor (CA) shared among members of the TG.[112,113] Factors that affect the reliability of triangulated conclusions are constantly under discussion. In one such discussion:

- Bettinger advocates the need to understand the coverage and accuracy of the trees contributing the common ancestor.
- Bettinger, and also Johnston, expect a matching segment to be sized such that it is probable it is identical-by-descent (IBD; see *IBD vs. IBS vs. IBC* on page 40 for more information).
- Bettinger expects quantitative analysis of total shared atDNA to fit anticipated ranges.
- Bettinger gives Bartlett's intermediate CA concept possible merit, while Johnston seems wholeheartedly in favor of declaring it required.
- Both Bartlett and Johnston believe that the matching segments will fit (without conflict) within the chromosome maps that have been established for the genotyped individuals.
- Johnston wishes to impose a number of other constraints including genotyped individuals knowing their chromosome crossover locations, and grandparent phasing.[114,115,116]

It is unlikely that all solutions will satisfy all demands. Genetic genealogists will need to be able to evaluate their solutions against a number of criteria and convey information about strengths and weaknesses inherent in their solutions.

Evaluating Matching Segments

Another type of quantitative analysis that is important to identifying matching atDNA segments as IBD is the analysis of the length/size of the matching segment. In one definition, IBD matches are characterized by their frequency (or rather their rarity)—the likelihood that an identical sequence in the same position could be observed in two separate individuals at the same time.[117] Segment size (given in centimorgans) becomes a proxy measure for segment frequency: the longer the shared sequence, the smaller the likelihood that it could be observed by chance in separate individuals. In fact, a centimorgan is defined in terms of likelihood.[118]

Much attention is given to whether a small half identical region (HIR)—often called a matching segment in the literature—can be used in making genealogical inferences (i.e., can be shown to be IBD).[119,120] Janzen published data about two genomes he studied extensively showing that nearly 80% of 5 cM HIRs could not be IBD.[121] Prairielad, an author in the Family Tree DNA™ forums, posted similar data from 14 genomes studied in his family.[122] In these instances, the small numbers of genomes involved leave unresolved questions about the statistical validity of their conclusions. Walden shared data from a private study of 9000 genomes, showing that even small, phased matches are largely IBS (not IBD).[123] This study has received criticism for its lack of disclosure and peer-review.[124] The relevance of small HIRs remains a critical topic; experts continue to advocate extreme caution when considering these regions.[125]

Quantitative evaluation of HIRs is also tied to the nature and quality of genotype data. With two measured allele values per SNP (and with possibly missing/uncalled values), there is a likelihood that a match will be declared where there should be no match—a false-positive. At times, this likelihood is quite high.[126] For this reason, long HIRs are required for confidence in IBD conclusions.[127] Even phasing (discussed below) cannot eliminate the possibility of false matching as HIRs get shorter. While thought leaders warn against trusting small HIRs, not all possible sources of error are included in their reasoning.[128,129]

Phasing

An atDNA genotype is a collection of SNP assay results. The SNPs examined are selected to show common variation—to highlight areas in the human genome where one is likely to find differences, not areas that are always the same.[130] SNPs are located at well-known sites (loci) in the genome. Because a human genome is diploid, every autosome SNP that is measured generates two values—one allele (base value) coming from the paternal haploid and one from the maternal haploid. There is no mechanism in the assay process to distinguish which of the two base values reported for a given SNP is paternal or maternal—but one surely is paternal and the other maternal. The process of assigning alleles to a parent-specific chromosome is called phasing.[131] Phased genotype data is very valuable to many aspects of genetic research.[132] For the genetic genealogist, phasing greatly assists in the identification of IBD matches—or perhaps more accurately, the elimination of a great number of matches that are IBC.[133,134] Phasing does not ensure that all remaining matches are IBD, as some would like to believe.[135,136] Using his own genome, Janzen reports that only 23% of his "matches" remained when using phased data for comparison at a 7 cM threshold.[137]

Chromosome Mapping

Genetic genealogists undertake chromosome mapping to associate specific matching atDNA segments (HIRs) with specific ancestors.[138] It can also be used to associate specific HIRs with ethnic populations.[139] By way of example, Janzen has mapped 95% of his mother's genome as to whether it came from her father or her mother, but he has mapped only 30% of his mother's genome to specific grandparents.[140] Because triangulation associates HIRs with known ancestors, chromosome mapping and triangulation are integrally connected.[141,142] When a matching segment does not fit properly within a mapped region of a chromosome, it can signal that the matching segment is IBS.[143,144] In the genetic genealogical community, Janzen's use of this technique is well known and his expertise often sought.[145] Janzen, sometimes collaborating with others, has been forthcoming in explaining his processes.[146,147] Others have provided tools for representing maps. Kitty Cooper has built and published several on-line tools to assist in the graphical representation of chromosome maps.[148] Griffiths has provided several spreadsheets for representing chromosome maps.[149]

RESEARCH PROCESS

This text's focus on methodology was born out of the author's own desire to make use of his genotypes for genealogical purposes. His first genotype (one representing his aging grandmother) was made available April 6th, 2014.[150] His own genotype was available starting February 8th, 2015.[151] Yet, the author had not made use of these important resources. He had examined the information provided by the genotype services, had read various published materials, and had attended numerous presentations about genetic genealogy. Even so, there was no clear place to start or path forward.

The author was very obviously not alone in this feeling. Most of the people sitting with him in presentations seemed to face a similar dilemma. The hands of most present would be up when asked if they had been genotyped, but most hands were down when asked about doing more with their results than reviewing their ethnicity estimates. People in the genealogical community are very interested in using their genomes to further their genealogy but, from novice to expert, genealogists seem to feel confusion and a lack direction as they seek to make use of their genotypes.

When proposing a solution to a cousin's "brick wall" (detailed in *Appendix F*), the cousin was unwilling to believe the solution as initially proposed because the literature and materials common in the community do not address the nuances that make the solution viable. As the author sought to understand the necessary details and to address the concerns raised, it became apparent that the elements important to the process were not well identified and often only discussed in isolation. Even after months of exploring genetic genealogical literature, a vision of what the ultimate goals ought to be and a path to reach those goals had not emerged. The needed roadmap was missing. A methodological framework was needed.

The framework presented in this text is grounded in the genealogical research process and the genealogical proof standard. The author first encountered these topics as evangelized by Elizabeth Shown Mills at a National Genealogical Society (NGS) conference in 2010.[152] He has continued exploring these topics in subsequent years—in Mills' book *Evidence Explained*, in conference talks, and in other literature (see the section titled *Proof* above). In 2013, the author participated in the Salt Lake Institute of Genealogy (SLIG), completing the *Advanced*

Genealogical Methods course (taught by Thomas W. Jones)—a deep dive into genealogical research methodology.[153] It is Jones' explanations of methodology that have most resonated with the author and that are the foundation of the genealogical research process and genealogical proof concepts presented herein. Subsequent collaboration in the development of the GEDCOM X project (including collaboration with Mr. Jones) has served to further deepen and refine the author's knowledge and understanding of these concepts.[154]

The author spent time reading a wide variety of blogs and other published materials seeking to understand the use of atDNA for genealogical purposes. The author looked for methodological material, and then for answers to specific questions relative to his research issues. Methodological material tended to be step-by-step directions (the *how* without much content as to *why*). The author also read a wide variety of material published in academia (e.g., Speed) and by expert members from the genetic genealogy community (e.g., Bettinger, Janzen). Addressing the concerns raised in the case detailed in *Appendix F* was particularly important; this focused the research and served to highlight the most important information.

As the author explored the various concepts strewn around the landscape of the genetic genealogical literature and tried to synthesize these concepts into an integrated whole, he began to see how the fundamental elements could be mapped into the conceptual framework and methods of the genealogical research process. The needed methodological framework already existed! Mapping the concepts and processes from the genetic genealogical community into that context caused a genetic genealogical methodology to emerge.

The information (i.e., genotypes and compiled genealogies) used in this text has been gathered using methods typical of genetic genealogists (e.g., vender-specific messaging, email, phone calls, published genealogies, etc.). The author sought to experience genetic genealogy in the way it would be experienced by any other researcher. Because genotyping services are not providing all the tools required to execute the methodology detailed herein, the author focused on genotypes available in GEDmatch (a service that does provide tools sufficient to execute this methodology). In many cases, genotype administrators were willing to make their genotypes available in GEDmatch when asked.

The author worked with genotypes that match the familial genotypes that he administers. AncestryDNA™ had matched the author with about 3600 genotypes, and 23andMe™ with about 900, at the time he began work on this text.[155] Contacting the administrators for each of these genotypes in a systematic way was never a priority for this project. Rather, the author was opportunistic in his approach—interacting first with genealogists actively researching their connection with the genotypes administered by the author, then seeking the cooperation of the administrators of related genotypes. The author also sought the cooperation of close cousins that were already genotyped.

Given a cooperating genotype administrator, the first goals were to perform any testing that could eliminate the match as IBC and to establish the existence of a common ancestor. If the genotype was not likely IBC and if a common ancestor was identified, other genotypes that shared this same matching segment were identified—i.e., the “triangulated group” for the matching segment was identified. At this point, an analysis of genotype independence was performed to identify any family groups within the triangulated group. Using compiled genealogies that represented the individuals (or family groups) in the triangulated group, common ancestors (if any) were identified with these additional genotypes.

Over the course of this research, the correspondence with genotype administrators surpassed 700 messages. A common ancestor was identified with more than 50 genotypes.

To anonymize the data presented in this text, genotypes have been assigned an identifier specific to this text. The identifier is in the form of GT000—“GT” for genotype, followed by three randomly generated digits. The mapping of these identifiers to vender-specific identifiers will be submitted with the text, but not published. Except in one case, the actual details of persons (living or deceased) have been omitted from lineages, etc. The exception is a case where a “brick wall” necessitated a discussion of the genealogy of a few specific ancestors (*Appendix F*). In this case, only identifying details of deceased individuals are included.

METHODOLOGY

GENEALOGICAL RESEARCH PRINCIPLES AND PRACTICE

A *conclusion* is a *hypothesis* that passes scrutiny. To form a *hypothesis*, the researcher correlates at least two independent instances of *evidence*. *Evidence* is the result of analyzing and correlating *information* to answer a question. *Information* comes from a *source*.

The genealogical research process begins with question asking. The question should not be too broad (where more than one correct answer is possible) nor too narrow (where an answer is not possible because the sources available are not capable of producing an answer).[156]

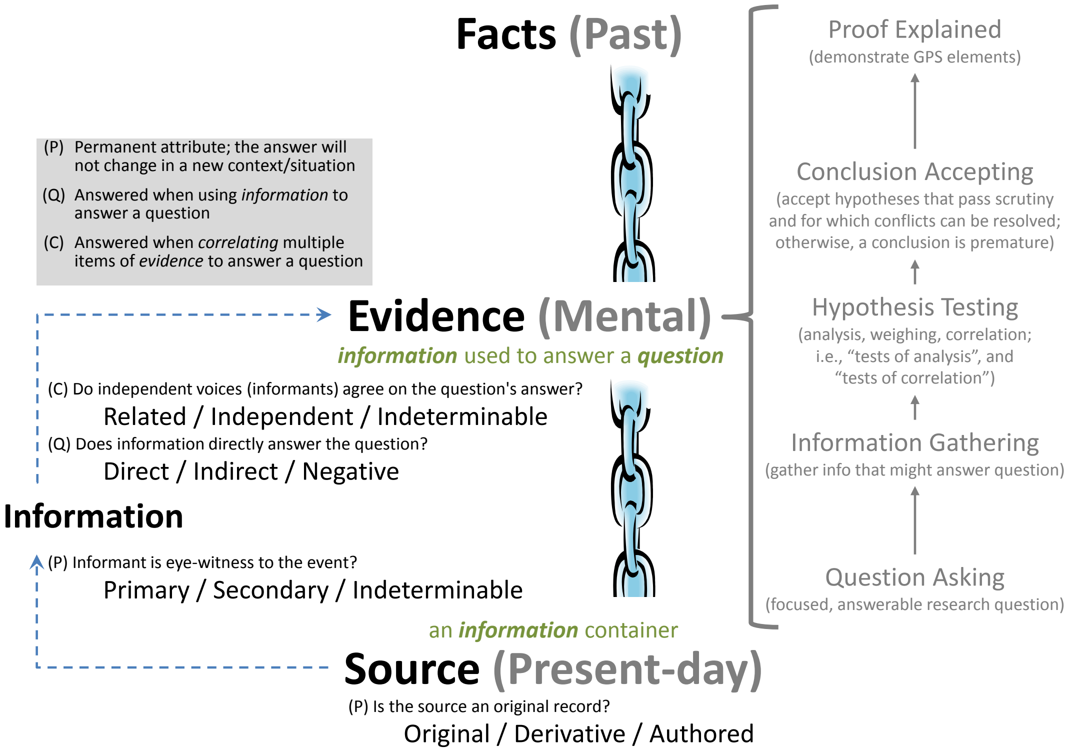
Armed with a proper question, the genealogist shifts to *information* gathering. A thorough search is planned, identifying all *sources* that might reasonably contain *information* that could answer the question.[157] Each *source* is examined and the results (positive or negative) are noted. If the examination of a *source* leads to additional *sources*, the genealogist repeats the process for each additional *source*.

The genealogist analyses all relevant *information* for answers to their research question—each tentative answer an item of *evidence*.[158] A minimum of two independent items of *evidence*—two independent answers to the research question—must correlate (agree) to form a credible *hypothesis*.[159] All relevant items of evidence must be synthesized to form an integrated whole.

As the researcher synthesizes the relevant evidence to the answer the research question, the *evidence* forming their answer is tested—the tentative answer becoming a *hypothesis*.[160] *Hypothesis* testing considers whether items of *evidence* are suitable and able to answer the research question. Each building block used to establish the *hypothesis* (each *source*, *information* item, and/or item of *evidence*) is examined. Tests of analysis and tests of correlation are considered. Tests of analysis evaluate the likelihood a building block is what it was assumed to be. Tests of correlation examine whether independent items of *evidence* agree.

A *hypothesis* is accepted as a *conclusion* when it passes all applicable tests of analysis and correlation.[161] In the GPS, a proven *conclusion* requires a written *conclusion*.[162]

The genealogical research process is summarized in Figure 4 as follows: *Sources* (people and artifacts available in our present day) are interrogated for *information* that, through mental analysis and manipulation—i.e., the genealogical research process of question asking, information gathering, hypothesis testing, conclusion accepting and proof explained—are transformed into *evidence* that is used to answer questions about the facts of the past.


Facts (Past)
(P) Permanent attribute; the answer will not change in a new context/situation
(Q) Answered when using *information* to answer a question
(C) Answered when *correlating* multiple items of *evidence* to answer a question
Evidence (Mental)
information used to answer a *question*
(C) Do independent voices (informants) agree on the question's answer?
Related / Independent / Indeterminable
(Q) Does information directly answer the question?
Direct / Indirect / Negative
Information
(P) Informant is eye-witness to the event?
Primary / Secondary / Indeterminable
an *information* container
Source (Present-day)
(P) Is the source an original record?
Original / Derivative / Authored
Proof Explained
(demonstrate GPS elements)
Conclusion Accepting
(accept hypotheses that pass scrutiny and for which conflicts can be resolved; otherwise, a conclusion is premature)
Hypothesis Testing
(analysis, weighing, correlation; i.e., "tests of analysis", and "tests of correlation")
Information Gathering
(gather info that might answer question)
Question Asking
(focused, answerable research question)

Adapted with permission from Thomas W. Jones, "Schematic of Genealogical Methodology," figure In course material for Advanced Genealogical Methods (Salt Lake Institute of Genealogy, 2013), p. 6.
Also from Thomas W. Jones, "Systematic Genealogical Research's Five Phases" in "Planning Efficient and Effective Research: A Case Study" handout for evening session talk of the same name (Salt Lake Institute of Genealogy, 2013), p. 1.

Figure 4: A representation of the genealogical research process.[163]

Methods used in genetic genealogy should (and do) fit into the genealogical research process. As genetic genealogical practice is considered in what follows, the methodology is grounded in the genealogical research process:

1. question asking
2. information gathering
3. hypothesis testing
4. conclusion accepting
5. and proof explained

Considering the methodology in this way integrates genetic genealogical practice within the broader genealogical research practice.

Question Asking

A genetic genealogist, like any researcher, needs focused, answerable questions—questions suited to the unique characteristics and capabilities inherent in the human genome. Yet would-be genetic genealogists are sometimes unrealistic in their expectations. It is appropriate, therefore, to consider the possibilities and limitations before one engages in a genetic genealogy research project.

Genetic genealogists have found genotyping useful for the following:

- finding cousins
- confirming relationships (e.g., when paper trails are weak)
- disproving relationships
- breaking through brick walls
    - results can suggest relationships where none were previously known
    - results can reveal/correct misattributed paternity
- finding biological parents (for someone living, or for a deceased ancestor)
- exploring ethnic origins
- exploring deep ancestry (requires mtDNA or yDNA haplotypes)

New genetic genealogists often feel overwhelmed as they attempt to make sense of the information provided by genotyping companies. They express surprise at the amount of work and learning required to get the expected benefits. Yet, these benefits are real and achievable in many instances.

Information Gathering

With a research question in hand that may be answerable using atDNA, the genetic genealogist takes up the information gathering aspects of the research process.

Core Sources

Information is found in and gathered from *sources*.

In genetic genealogy, the primary sources of genetic *information* are human subjects—oneself, parents, siblings, children, cousins, grandparents, etc. These primary *sources* are living records—each unique (except in cases of monozygotic twins, triplets, etc.), and each with information about several generations of ancestors. Access to the *information* in these records comes via genotyping. Genotyping requires that one submit biological material to be assayed. Current direct-to-consumer genotyping services request a saliva sample or cheek swab to collect the necessary biological material.

Genetic *information* is necessarily interpreted in the context of pedigrees when used for genealogical purposes. Compiled genealogies become critical *sources* of pedigree *information* as research proceeds. These *sources* are authored works—presenting many *conclusions*, possibly containing information available nowhere else, often based on primary *information* from original *sources*. These *sources* (compiled genealogies) may be susceptible to author bias or mistakes made in interpretation. It is not practical for genetic genealogists to personally create and curate all of the pedigrees necessary to unlock the *information* available in these genomic records. Collaboration and sharing are critical to genetic genealogy success.

CORE INFORMATION

The core *information* item from a human *source* is their genotype. The core *information* item from a compiled genealogy is a pedigree extract that details a person's lineal relationship to a common ancestor.

GENOTYPES

Genotyping studies the genetic makeup of individuals by examining the genetic variants they possess.[164] Genotyping differs from genome sequencing (which attempts to measure every base value) in that it measures a small subset of alleles selected for their likelihood of variation.[165] The resulting data, as an aggregate entity, is called a *genotype*. In many research situations, a genotype stands as a proxy for an individual's entire genome.

In a typical atDNA genotype, the proxy data represents only about 700,000 SNPs of more than 3 billion sequenceable loci that make up a human genome—representing less than 1/5,000th of the total information that could have been measured.[166,167]

Despite its clearly sparse representation, a genotype can effectively represent the entire genome because an individual human genome is 99.9% identical to all other human genomes.[168] Though sparse, genetic genealogists still consider atDNA genotype data in a sequenced way.

With the raw SNP data, genotyping companies provide at least two other *information* products: a list of DNA matches (persons identified as having genotypes with identical regions in common with the given genotype), and ethnicity estimates.

What is an identical region? Who can I expect to find in my match list? How can an ethnicity estimate be useful? Why do ethnicity estimates vary from company to company and from tool to tool for the same genotype? Exploring these topics further will facilitate the use and prevent the misuse of this important *information*.

IBD vs. IBS vs. IBC

Genetic genealogy has its foundation in the principles of genetic inheritance. Genetic genealogy becomes possible when two genomes have one or more identical regions—matching segments—that are identical-by-descent (IBD). In an atDNA genotype, a segment is the allele values from a specific region of a single haplotype (chromosome). Two segments are IBD if they are identical and if they were inherited from a *recent* common ancestor.

A segment is a region of a single chromosome—its allele values all a part of a single haplotype, and each haplotype representing a single chromosome (e.g., Chr1). Because each chromosome is diploid, each SNP assay results in two measured base values—one from the father, and one from the mother. The assay is not haploid-specific for parent origin; it does not report which base value came from which parent.

Assayed:	C/A	A/C	C/C	A/T	G/T	C/C	C/A	G/C	G/T	G/A	A/T	A/G
Father:	A	A	C	T	G	C	C	C	T	G	T	G
Mother	C	C	C	A	T	C	A	G	G	A	A	A

Figure 5: Un-phased haplotype is not haploid-specific.

Now consider asking whether two haplotypes are identical. Haplotypes are identical if, at a given SNP, they have at least one allele in common.

Identical-By-Chance (a false match)												
Assayed:	A/G	A/T	C/A	T/C	G/C	A/C	C/T	C/T	C/T	G/C	T/G	G/C
Father:	A	T	C	C	G	A	C	T	T	C	T	C
Mother	G	A	A	T	C	C	T	C	C	G	G	G

Identical-By-State (a paternal match; potentially IBD)												
Assayed:	C/A	C/A	C/C	A/T	T/G	C/C	C/G	A/C	A/T	T/G	G/T	G/C
Father	C	C	C	A	T	C	G	A	A	T	G	C
Mother:	A	A	C	T	G	C	C	C	T	G	T	G

Figure 6: Two genotypes that match with the Figure 5 genotype: the first is identical-by-chance; the second is at least identical-by-state and matches the paternal haploid.

Matching algorithms will declare both haplotypes in Figure 6 to match the haplotype in Figure 5. However, the first match has nothing to do with inheritance. It is an artefact of the data and its lack of haploid-specificity. The match is declared using a mix of alleles from both the paternal and maternal haploids. Because the matching algorithm does not know from which parent (haploid chromosome) the allele originated, it ends up declaring a match that is not biologically meaningful—a false-positive. The author designates matches of this type as identical-by-chance (IBC).

The second haplotype in Figure 6 has an allele sequence along its maternal haploid that matches the sequence along the paternal haploid in the Figure 5 haplotype. This match is, at a minimum, identical-by-state (IBS). It could be IBD, but this designation requires that other factors be considered.

When comparing two sets of unphased allele values from two haplotypes, if an allele from each set is identical, the data are said to be half-identical at that location. If a shared region is composed of half-identical alleles, it is said to be a half-identical region (HIR). HIRs may be IBC, IBS or IBD. When one considers the HIR that is the result of comparing the paternal haplotype from Figure 5 with the first haplotype in Figure 6, one can see that the HIR in Figure 6 is IBC because the matching alleles from Figure 6 are a mix of alleles from both parents.

It is possible, especially in the case of siblings, that regions are identical simultaneously on both chromosomes. Such a region is designated a fully-identical-region (FIR). The result of comparing the allele values from the Figure 5 haplotype with the allele values in the second haplotype in Figure 6 is an FIR across the first half of the region, and an HIR across the second half.

The genetic genealogy community generally does not differentiate the terms IBS and IBC, using them interchangeably.[169,170] A few in the community seem to use IBC in the manner it has been employed in this text.[171,172] In at least one instance, this author found a use of IBC that may have been better characterized as IBS.[173] It is true that the above IBC match is IBS (making IBC matches a subset of all IBS matches). However, the author believes that it is useful to consider matches that are a mash-up of the alleles from both parents separately from other types of IBS matches.

By definition, every base pair (and, therefore, every allele) in one's genome was inherited from an ancestor (except where mutation has modified a base pair after inheritance has played its role). This is why the notion of *recent* is part of the IBD definition. A timeframe for *recent* is not formally defined, but it would be difficult to classify an ancestor as *recent* if the common ancestor is outside of "genealogical time" (the timeframe in which it is reasonable to expect it is possible to document the genealogy of one's ancestors). The *recent* designation is also important in that we need to be able to identify the common ancestor (in time and space, and our relationship to them) to show inheritance. However, Speed and Balding have shown the possibility of sizeable IBD segments from ancestors more than 20 generations from the subject—well outside the possibility of documented genealogies for most.[174] Sometimes, shared atDNA is identical for historical reasons (e.g., endogamy or ethnicity)—IBS, but no longer distinctly inherited from a single common ancestor—separating *recent* by contrast from *historical* atDNA. Therefore, we might define *recent* by contrast as *not ancient*.

Match Lists

In addition to the *information* from SNP assays, genotyping services that enable genetic genealogy provide a list of matches—other genotypes in their database that share regions that are IBS with regions in one's own genotype. Matches in these lists are generally assigned an estimated relationship (e.g., "2nd Cousin – 3rd Cousin" or "5th Cousin – Remote Cousin") and grouped according to these estimates. Typically, these lists are ordered by the amount of IBS sharing in centimorgans (with those sharing the most listed first).

Unless one has recruited (or been recruited by) close relatives, it is not typical to find many close relatives in their list of matches. The bulk of one's matches will be (predicted) distant relatives. This is because the more likely sharing is IBD, the rarer (lower frequency) the matches will be. (See also *Matching Segment* Size on p. 64.)

Consider also that the number of distant cousins one has is much greater than the number of close cousins. Henn et al estimated one might have 4,700 fifth cousins, 23,000 sixth cousins, and 120,000 seventh cousins.[175] As Steve Mount puts it: "…if you have many more distant cousins, as would be expected if your ancestors had large families, then someone who shares a single IBD segment is more likely to be a distant cousin, because you have so many more distant cousins."[176]

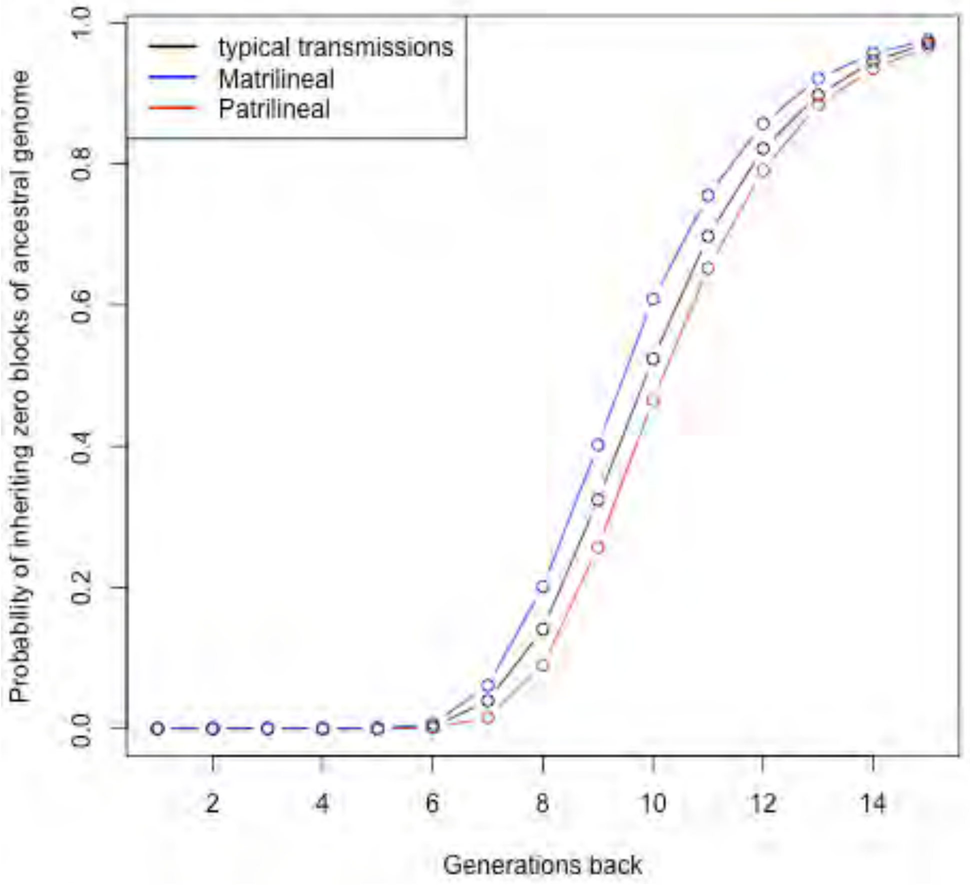


Figure 7: Probability of inheriting zero (large) blocks of ancestral atDNA.[177]

How many generations of ancestors on all of one's ancestral lines are detectable in one's atDNA? The genetic genealogy community's answer has been five generations.[178] Donnelly (1983) and Jostins (2009) are cited in support of this notion.[179,180] However, in 2013, Jostins explicitly deferred to the work of Graham Coop. Coop predicts that the likelihood of having zero detectable atDNA from an ancestor will not rise significantly until the 8th generation—giving researchers a high likelihood of carrying detectable atDNA regions for nearly all of their 7th generation ancestors (see Figure 7). Beyond the 7th generation, the number of genetic ancestors (persons with whom one shares atDNA) is no longer the same as the number of genealogical ancestors (persons in one's pedigree). Bartlett affectionately calls a fan-chart plot of genetic ancestors a porcupine chart (see Figure 8).[181]

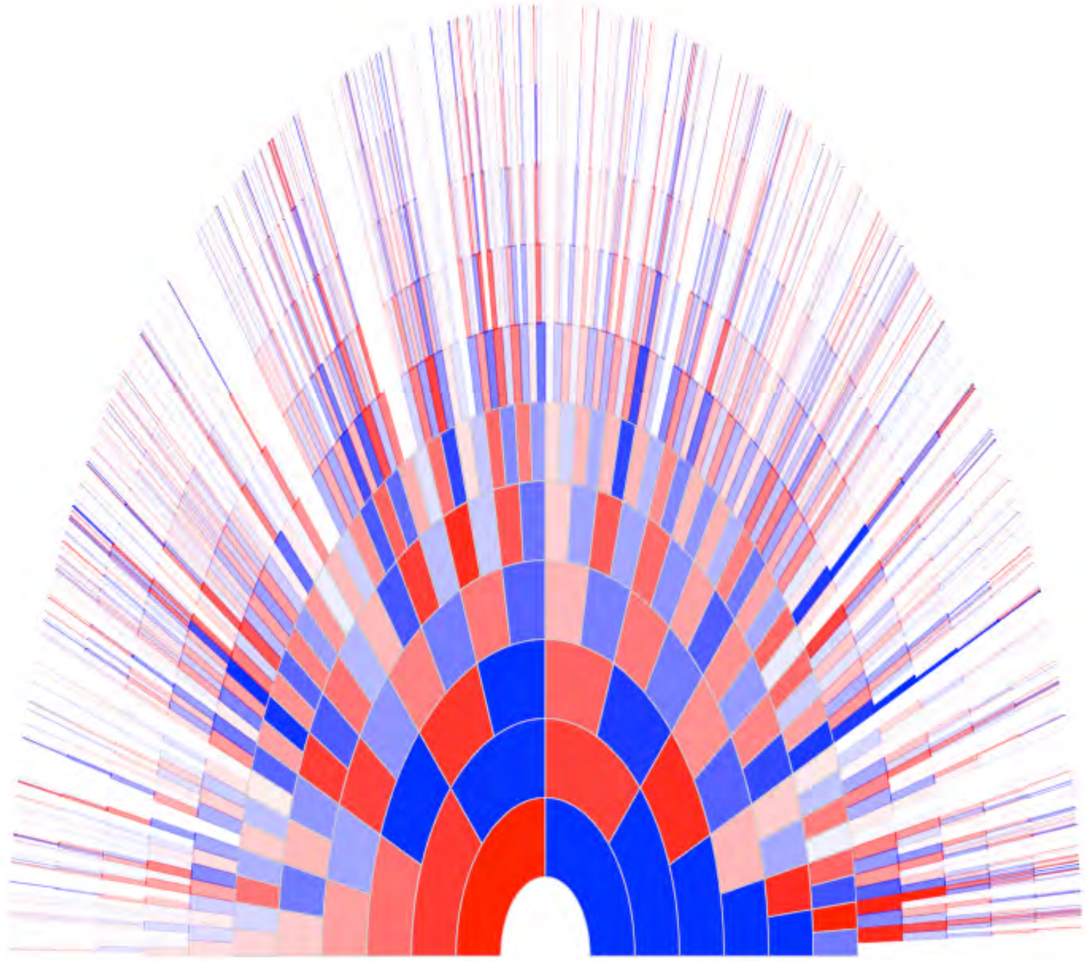


Figure 8: A fan-chart plot of 11 generations of genetic ancestors (simulated)—a so-called porcupine chart.[182]

Ethnicity Estimates

Along with assayed results and match lists, genotyping services provide *information* about ethnicity. Ethnicity estimates involve comparing a genotype to a reference population—a set of genotypes selected to represent a given ethnicity.[183] Algorithms seek a genotype's "best fit" with the collected reference populations.[184]

Each genotyping service's ethnicity reports are different, and individual results will differ between services. Hence, it has been said that estimating ethnicity is "part interpretive art and part science."[185]

Many express frustrations because their estimates do not correlate with what they know of their ancestry. The process involves comparing genotypes; it is susceptible to all the problems of such comparisons (e.g., being IBC). Reference populations may not actually represent the actual population of one's ancestor. One may no longer carry atDNA from the ancestor that was part of that population. Roberta Estes' article *Ethnicity Testing – A Conundrum* explains many of the pitfalls associated with ethnicity *information*.[186]

Ethnicity *information* can be used for genetic genealogical purposes, but it requires knowing which segment(s) are associated with the ethnicity of interest. Except for 23andMe™, genotyping companies do not provide this information. Triangulation can also be used to identify the segments of interest. The GEDmatch *Admixture/Oracle Population Search Utility* has reference populations that can be used to assist in this process.[187] Estes' article touches briefly on these methods.

LINEAGES

Genotypes cannot be converted into genealogical *information* without knowing the individual *source* subjects and their pedigrees. The *information* needed about these subjects is the relationship(s) (if any) that exists between them. Genetic genealogists seek the ancestors common among individuals—the point(s) at which two pedigrees intersect. Compiled genealogies are the *sources* that contain the *information* that is searched to identify these common ancestors and the lineages that link these *source* subjects to these common ancestors.

Ideally, the search for a common ancestor involves pedigrees that are complete (all ancestors identified) to at least the generation that includes the common ancestor, that are sourced with original records, and that exhibit only one ancestor in common. Practically speaking, there will be many incomplete pedigrees in various stages of development. Each compiler possesses different capabilities and different access to the genealogical records. Searching these pedigrees may lead the genetic genealogist to additional activities such as filling in pedigree gaps, looking at descendants, correcting erroneous relationships, etc. Searches may also reveal more than one common ancestor—the reason for desiring a complete pedigree. *Conclusions* are susceptible to error/conflict when the search for a common ancestor does not include complete pedigrees for all individuals involved.

Once the common ancestor has been identified and the strengths and weaknesses of the compiled genealogies have been characterized, researchers generally extract and present enough *information* to document lineages from each subject to the common ancestor.

DEVELOPING A RESEARCH PLAN

In genealogical research practice, developing a research plan is about enumerating the *sources* that are likely to contain *information* that can be used to answer a research question, then developing a plan to exhaustively search those *sources* to gather the *information* that answers the research question. Genetic genealogists answer questions like the following:

- Which persons, if genotyped, are likely to possess the atDNA needed to answer the research question?
- If these persons are genotyped, how can I access the genotypic and genealogical *information* necessary to answer the research question?
- If these persons have not been genotyped already, what can be done to obtain their genotypes?
- What can be done to prepare to make use of these *sources* when they become available?

IDENTIFYING GENOTYPE SOURCES

Genetic genealogy might be easier if everyone was genotyped. This is not a realistic option. Who, then, are the priorities?

If research questions address distant generations, the highest priority genotypes are the genotypes separated from these distant generations by the fewest meioses. If one is looking for the parents of a 3rd great-grandparent, it would be better to genotype one’s parent, aunt/uncle, or grandparent (who also share a relationship with that 3rd grandparent) than oneself. It may be that a great-uncle or a second cousin twice removed is available to be genotyped. Seeking the youngest child of the youngest child, etc., might yield a living descendent that could be genotyped with a smaller generational gap than one’s own close family.

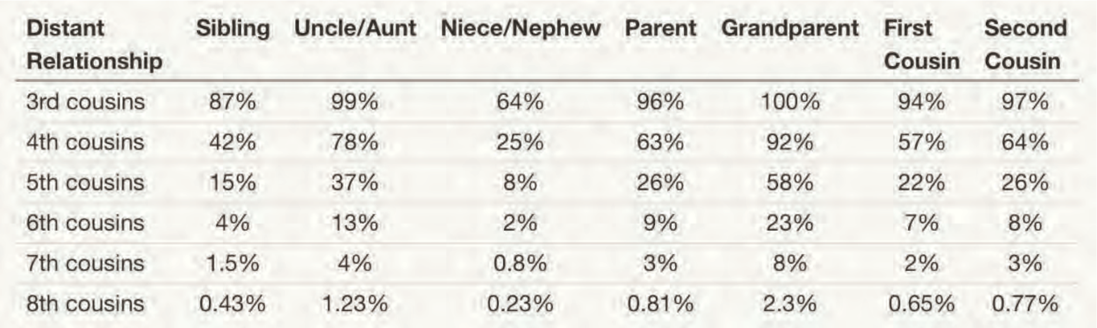
If one’s research interest is broad (not focused on a particular ancestor) but interest is high for distant generations, start with members of the oldest living generation (e.g., grandparents).

If seeking to access the genome of a parent not available for genotyping, gather the genotypes of their children.

In many cases (e.g., seeking biological parents), the persons with the needed genotypes are not known in advance. The only option is to wait for the right person(s) to become interested in exploring their genotype. There is plenty of reason to remain hopeful. Bartlett estimates that genotype databases are doubling every fourteen months.[188] Today’s missing genotype may be available next year.

Phased genotypes help eliminate matches with genotypes that are IBC. Parent genotypes are the easiest path to a phased genotype—making parent genotypes invaluable to researching with one’s own genotype.

For some, a dearth of matching genotypes inhibits research. Making one’s genotype available in all the major genotype databases (“fishing in all of the ponds”) could help in finding the needed matches. Figure 9 shows the value of genotyping other relatives to uncover genotypes that did not match one’s own. Some of the values in Figure 9 might be surprising; for instance, genotyping an aunt/uncle is more likely to reveal a new fourth cousin than testing a parent.



Distant Relationship	Sibling	Uncle/Aunt	Niece/Nephew	Parent	Grandparent	First Cousin	Second Cousin
3rd cousins	87%	99%	64%	96%	100%	94%	97%
4th cousins	42%	78%	25%	63%	92%	57%	64%
5th cousins	15%	37%	8%	26%	58%	22%	26%
6th cousins	4%	13%	2%	9%	23%	7%	8%
7th cousins	1.5%	4%	0.8%	3%	8%	2%	3%
8th cousins	0.43%	1.23%	0.23%	0.81%	2.3%	0.65%	0.77%

Figure 9: The likelihood that a genotype of a close relative from one side of the family will match a genotyped distant cousin from the same side of the family that does not match one’s own genotype.[189]

RECRUITING SOURCES

Genetic genealogy is collaborative—requiring lots of participation and sharing.

Access to the right genotypes often requires recruiting relatives to be genotyped (which often includes paying for the creation of those genotypes). Some will not be willing to participate. Others might be willing but will find the sampling mechanism to be difficult (e.g., older persons often have difficulty providing a saliva sample), necessitating a different sampling technique (e.g., a cheek swab) and different genotyping service. Recruiting will likely require education—i.e., helping relatives understand the ramifications of being genotyped, helping them understand the results expected, helping them interpret the results received, etc.

Genotyping services do not provide all the information and tools necessary to use genotypes genealogically. Genetic genealogists need to know the following for each region shared with another genotype: the chromosome on which it is located, the start and end positions, the number of SNPs, and the segment size. Some of the required genotype-to-genotype comparisons cannot be made—in some services (e.g., AncestryDNA™), not at all; in others (e.g., Family Tree DNA™ and 23andMe™), they can be managed but only with the cooperation of other matching individuals. Recruiting genotypes of interest to other services (e.g., GEDmatch) that provide the necessary tools and information will maximize genotype utility. Again, this can require education—i.e., assisting these individuals with the technical challenges of doing so, helping them to understand privacy issues, etc.

Genetic genealogists require collaboration in the form of shared genealogies—i.e., published (public or private) genealogies that give information about one's ancestors. Researchers need to be prepared to share the genealogy of their genotyped person-of-interest. They will be requesting access to and examining genealogies compiled by others. They may need to assist others in creating or expanding pedigrees.

PREPARING YOUR OWN FAMILY TREE

If one's genealogy is known (or knowable), collect and document that genealogy. How much of one's genealogy is needed? Speed and Balding's work suggests an ideal tree might include twenty generations of ancestors and fully populated descendant trees for each ancestor—an impossible standard.[190,191] While the reasons for these statements have yet to be discussed, the author wishes to note that no one has (or can even create) such an ideal family tree. The essential work, then, is to gather the knowable genealogy—both ancestors and their descendants.

Information Gathering Is On-going

Information gathering is an on-going effort. Genotyping services continue to expand their databases which results in continual expansion of genotype match lists. New matches mean new cousins and new opportunities to collaborate. It will be necessary to revisit match lists, triangulated groups, research *conclusions*, etc.—keeping information current, adding individuals to circles of collaboration, and updating research *conclusions* to incorporate new findings.

Hypothesis Testing

With a critical mass of *information* gathered (i.e., matches to be analyzed, and compiled genealogies relevant to those matches), the *information* can be used to answer genealogical questions.

Consider, again, the genealogical research process. *Information* becomes *evidence* as it is used to answer a research question. *Evidence* suggests a tentative answer to a research question. At least two independent pieces of *evidence* must be correlated to have a testable *hypothesis*. Each *hypothesis* must be scrutinized—i.e., tested. A *hypothesis* that stands up to scrutiny is accepted as a *conclusion*. Testing safeguards the researcher from erroneous *conclusions* and helps to shine a spotlight on the strengths and/or weaknesses of those *conclusions*.

Tentative Answers

In genetic genealogy, researchers work with two kinds of *information*—genome *information* and relationship *information*—to find tentative answers to research questions, these answers becoming *evidence* of genealogical relationships.

EVIDENCE FROM QUANTITATIVE INFORMATION

The first *information* item used as evidence of relationships is the total amount of atDNA shared between the two individuals—a result of comparing two genotypes. This total can be (and has been) used to estimate the relationship between two individuals. This estimate can be very good for close relationships but breaks down as relationships become more distant. The following two figures illustrate the use of this basic metric.

The first comes from Bettinger's *Shared cM Project* (Figure 10). Researchers report the total amount of shared atDNA between two individuals along with the relationship believed to exist between them. Bettinger classifies each relationship within a "degree" of relatedness. Outliers for a given "degree" most likely represent errors in the reported relationship.

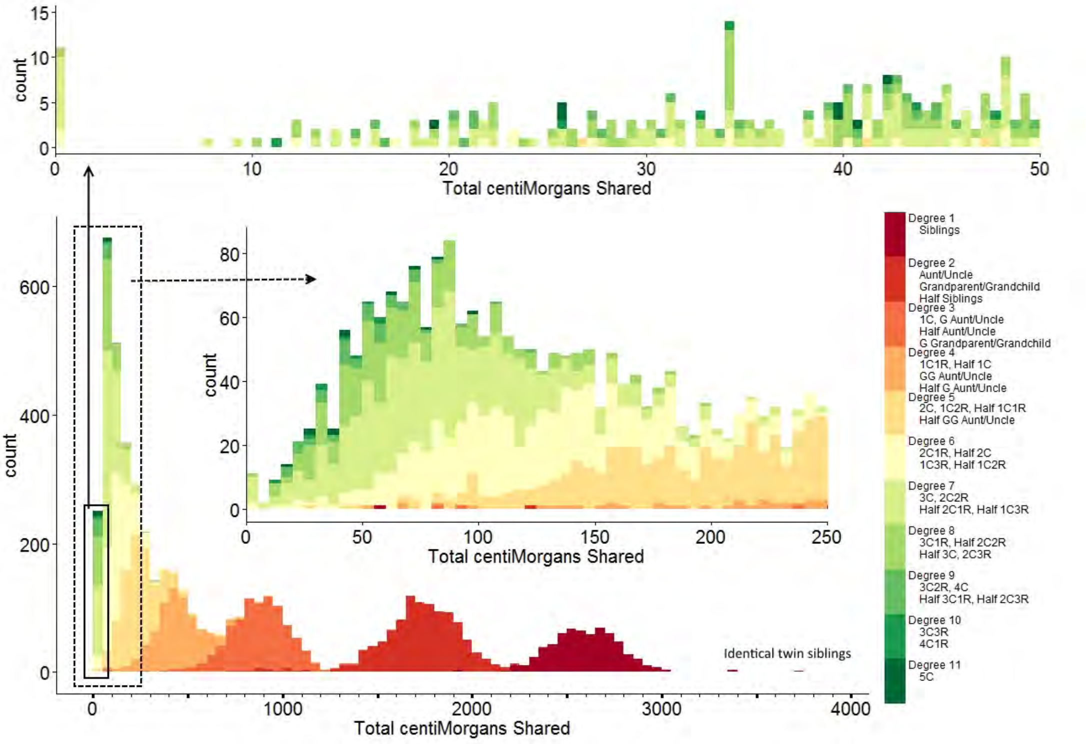


Figure 10: Distributions of shared cM by relationship type from the Shared cM Project.[192]

When considering these distributions, note that the distributions for the first several degrees on Bettinger's scale are very distinct from their neighboring degrees—with almost no overlap. If the total amount of shared atDNA between two individuals falls in the range of a first- or second-degree relationship, it would be illogical to argue that the relationship is somehow a third or fourth degree relationship. Contrast that to two individuals that share 100 cM of atDNA where it is much more difficult to discriminate relatedness. 100 cM is still in range with fifth-degree relationships, yet may also be in the ninth- or even tenth-degree ranges on Bettinger's scale.

Henn et al published a plot of data from a 23andMe™ data set (Figure 11).

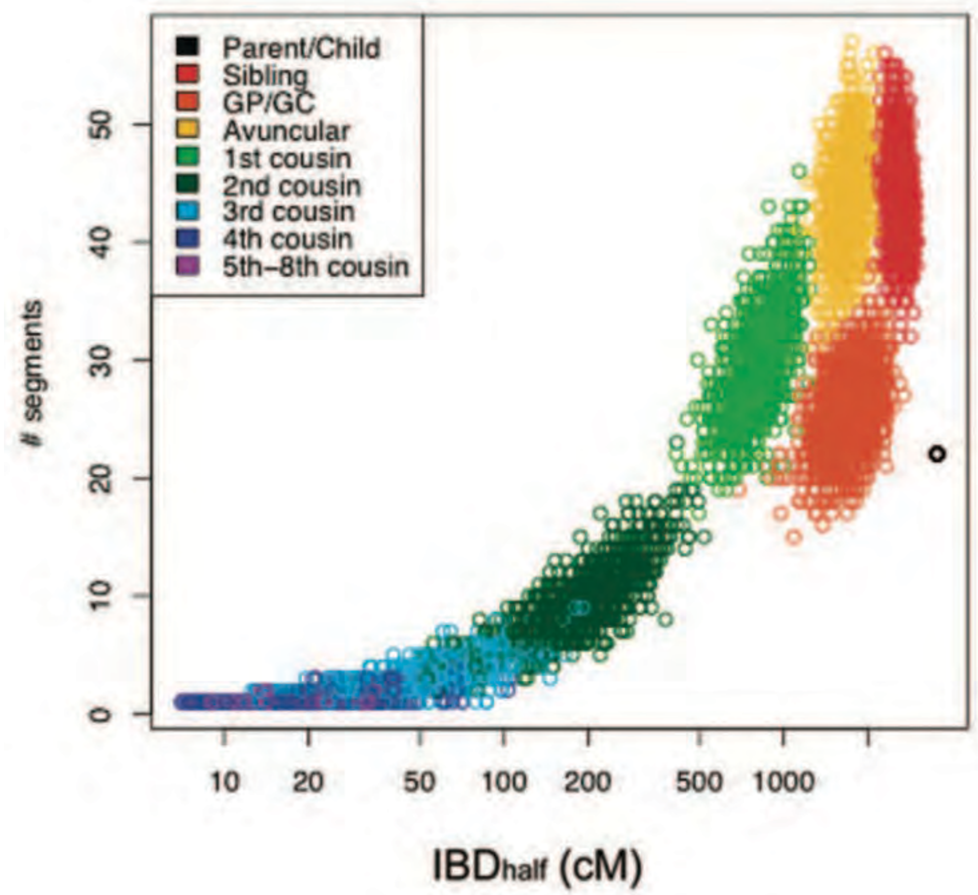


Figure 11: Total shared IBD$_{half}$ atDNA (x-axis) and number of segments shared IBD$_{half}$ (y-axis).[193]

Note the distinct data groupings for parent/child, sibling, grandparent/grandchild, avuncular and 1st cousin relationships in Figure 11. For these close relationships, each grouping seems to have little overlap with neighboring groupings. As groupings for distant relationships are considered, boundaries are no longer distinct. If one considers two genomes sharing five IBD segments totaling 100 cM, it is difficult to definitively identify the relationship shared between them.

As shown above, using the total amount of atDNA sharing as evidence of a particular relationship can be viable for close relationships. The data in Figure 11 suggests it would be safe to do so for first cousins and other closer relationships. The distributions from the *Shared cM Project* seems to suggest there may even be trouble with first cousin relationships and that the safe realm ends one degree closer than first cousin. The *Genetic Genealogy Standards* limit conclusions based solely on total sharing to first degree relationships.[194]

Consider the case of an adoptee—identified as an8181—searching for her birth parents via AncestryDNA™. Some point after receiving her initial results, AncestryDNA™ reported a match with a man designated M.G.[195] AncestryDNA™ reported that an8181 and M.G. shared 454 cM across 22 segments and estimated their relationship to be 1st-2nd cousins. Using the distributions in Figure 10, 454 cM of shared atDNA falls near the peak of Bettinger's Degree 4 relationship distribution. It would appear that the upper end of the Degree 5 distribution might overlap slightly with the Degree 4 distribution at 454 cM, but a Degree 4 relationship is clearly the highest probability answer.

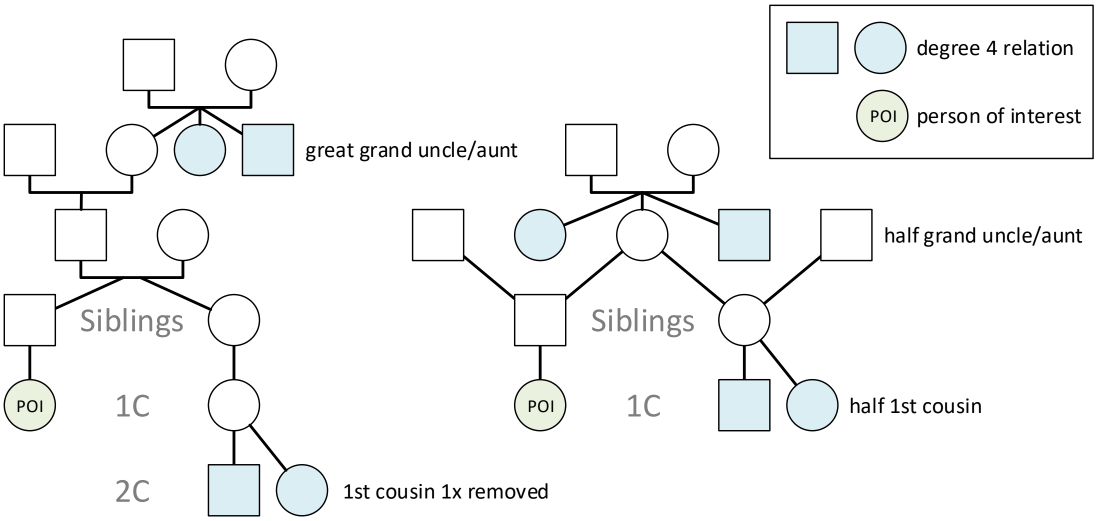


Figure 12: Relationships consistent with Bettinger's Degree 4 relatedness for a person of interest.[196]

Bettinger gives the following as Degree 4 relationships: 1st cousin 1x removed, half 1st cousin, great grand uncle/aunt, or half grand uncle/aunt.[197] When an8181's biological father was identified (see *Appendix A*), M.G. was shown to be her half 1st cousin—a perfect fit with a Degree 4 relationship.[198]

A NOTE ABOUT QUANTITIES

In genetics, segment size can be expressed in terms of physical length (typically in million base pairs—or Mbp) or in terms of genetic distance (centimorgans—cM). The total amount of sharing (the sum of the size of all the shared regions) is typically expressed in genetic distance (cM). When the genetic genealogist encounters information about size in the literature, tools, and reports that they use, sizes are generally expressed in cM—but not always.

While one might be tempted to consider the centimorgan to be a unit of length, it can be misleading when needing to consider an actual physical length. One can find many instances in the genetic genealogical community where a conversion is given between genetic distance (cM) and physical length (Mbp)—$1\ cM \cong 1\ Mbp$.[199] Given all of the matching segments m for GT999 where $10\ cM \geq size(m) < 11\ cM$, the smallest physical length encountered among these segments was 2.2 Mbp, and the largest was 39.9 Mbp.[200] So when considering heuristics that are given with genetic distances, seek to use sizes given in cM; and when considering heuristics given with physical lengths, seek to use sizes given in terms of base pairs.

EVIDENCE FROM TRIANGULATION

Genetic genealogists use triangulation to reliably attribute a matching segment (HIR) as coming from a particular ancestor. However, the literature explaining triangulation often fails to make plain the reasons why triangulation is effective. Pundits in the genetic genealogy community do not seem to agree broadly as to the elements required for a triangulated *conclusion*. Moreover, the community does not seem to apply uniform methods and criteria when evaluating the strengths and/or weaknesses of a triangulated conclusion. These deficiencies can be addressed if triangulation is decomposed into its fundamental building blocks and re-presented in the context of the genealogical research process.

AN AXIOM AND A THEOREM

Triangulation is born out of the following principle:

> A segment of atDNA shared by two individuals and received from a common ancestor is IBD.

This definition can be restated as follows:

> A segment shared by two individuals that is IBD was received from a common ancestor.

This principle is axiomatic. It is an axiom because it is true by definition. In mathematics, axioms are the premise for all further reasoning. This axiom is the cornerstone upon which genetic genealogists build—the basis for further reasoning—as they work to deduce identities and relationships from the genetic *information* related to their research interest.

The axiom above leads to the following triangulation theorem:

$$m \wedge a \wedge t \rightarrow c$$

where

m: a matching segment of atDNA, that might be IBD, that is shared by the individuals being considered

a: a recent ancestor, believed to be unique (the only one), that is shared in common among the individuals being considered

t: testing that supports the IBD and uniqueness suppositions in m and a (i.e., that supports a conclusion that the matching segment m was received IBD by the individuals being considered from the recent common ancestor a)

c: the contributor of the matching segment m for the individuals being considered was the common ancestor a.

In prose, it reads:

If m and a and t, then c.

In other words, the triangulation theorem can be expressed as follows:

If, for the individuals being considered, there is matching segment of atDNA (m) that might be IBD and a recent ancestor (a), believed to be uniquely common among them, and if testing (t) supports that the segment shared among them is IBD from the specific common ancestor, we can deduce that the recent common ancestor was the contributor (c) of the identified matching segment.

THE FUNDAMENTAL BUILDING BLOCK OF TRIANGULATION

The question that is being asked when triangulating is this: Can the presence of a shared segment (HIR) between two genotyped *source* individuals be attributed to a particular ancestor of those persons?

To answer this question, the genetic genealogist seeks *information* to answer two intermediate questions:

1. Does the *information* that resulted from genotyping the two *source* individuals substantially match (i.e., in an IBD-consistent manner) over the sequence identified by the segment in question?
2. Is there *information* in one or more compiled genealogies to suggest that the two *source* individuals share a single common ancestor?

If the answer to both of these intermediate questions is affirmative, there is *evidence* that the atDNA in common between the identified *source* individuals was contributed by the ancestor in common between them.

Notice that m and a in the triangulation theorem are answers to these intermediate questions. If the answers to these questions, together, are a piece of evidence that a particular ancestor contributed a particular segment (HIR) to one's genome, it is useful to refactor the triangulation theorem as follows:

$$\mathcal{E} \wedge t \to c$$

where

m, a, t, and c: as specified in the theorem above

$\mathcal{E}$: the conjunction of m and a—that is: $m \wedge a \to \mathcal{E}$

In prose, $m \wedge a \to \mathcal{E}$ says that if there is *information* indicating a shared match (m), and *information* documenting a common ancestor (a), there is *evidence* ($\mathcal{E}$)—a tentative answer—that the common ancestor (a) could be the contributor (c) of the shared match (m).

Often, the common ancestor between two donors is actually given as two ancestors—a father and a mother. Biology dictates that only one person in the couple could have actually contributed the matching segment (HIR) because an IBD segment is part of a haploid chromosome, and a haploid chromosome is received from only one parent.

In prose, $\mathcal{E} \wedge t \to c$ says that if the *evidence* ($\mathcal{E}$) withstands testing (t), there is a tentative answer suggesting that (c) is the contributor.

Considered as an entity, $\mathcal{E}$ is the fundamental building block in triangulation. A single instance of $\mathcal{E}$ is an item of *evidence* that becomes a tentative answer to the triangulation question. If two (or more) items of *evidence* (instances of $\mathcal{E}$) give the same answer (i.e., the items correlate), the genetic genealogist has a hypothesis that can be tested.

IDENTIFYING TRIANGULATION BUILDING BLOCKS

Given $\mathcal{E}$ as the fundamental building block and given that $\mathcal{E}$ is the result of m and a, an examination of m and a is merited.

ISOLATING m

m is the result of comparing two genotypes. In its simplest form, two individuals—considered *sources* in the context of the genealogical research process—submit biological material for genotyping. The genotype gives *information* about selected

SNPs dispersed throughout the autosomes. Because humans are diploid (each somatic cell containing two complete copies of the human haploid genome), a SNP measurement yields two allele values—one allele from each haploid copy. In the case of a SNP, each allele represents a nucleotide in a base pair; the second nucleotide in the base pair can be inferred from the value of the first—each nucleotide having a known complimentary base. Thus, the genotype is a collection of SNP allele values, each representing a nucleotide in a base pair at a known location on a known chromosome.

Matches (m) are determined by comparing two genotypes to discover matching segments (SNPs in sequence along a haploid chromosome) that might be IBD. Currently, all companies that offer a genome sequencing service also provide a list of genotypes that match the given genotype. Knowing that a match exists is not sufficient. A precise description of the match is required for triangulation. It is necessary to know:

- the chromosome that contains the segment
- the segment begin and end locations
- the size of the match (in cM)
- the number of SNPs present in the segment.

Genotyping companies are not consistent in describing matches. Vendors do not universally provide the necessary elements described above. To overcome this deficiency, genetic genealogists must persuade matches to add their genotypes into other databases—databases with tools sufficient to fully characterize each match.

Comparing two genotypes to identify shared regions is a form of *information* correlation. It transforms *information* about allele values into *information* about shared regions.

Isolating a

a comes from the *information* available in compiled genealogies. In the context of triangulation, compiled genealogies take on the role of *sources* in the genealogical research process. These genealogies are most useful when the persons and relationships contained in them are, themselves, documented using the GPS. A lineage extracted from such a genealogy showing an individual's relationship to the common ancestor is the *information* sought from these *source* genealogies.

A common ancestor is identified by consulting one or more genealogies and identifying the lineages within those genealogies that show how individuals share a biological relationship with a single ancestor (or ancestral couple). A common ancestor is “theoretical” in that the biological relationship is presumed and susceptible to question. In this text, a statement that a common ancestor exists is a statement that lineages exist that show this shared relationship to a presumed biological ancestor (or ancestral couple). Each such statement must be analyzed, evaluated, and weighed on its merits.

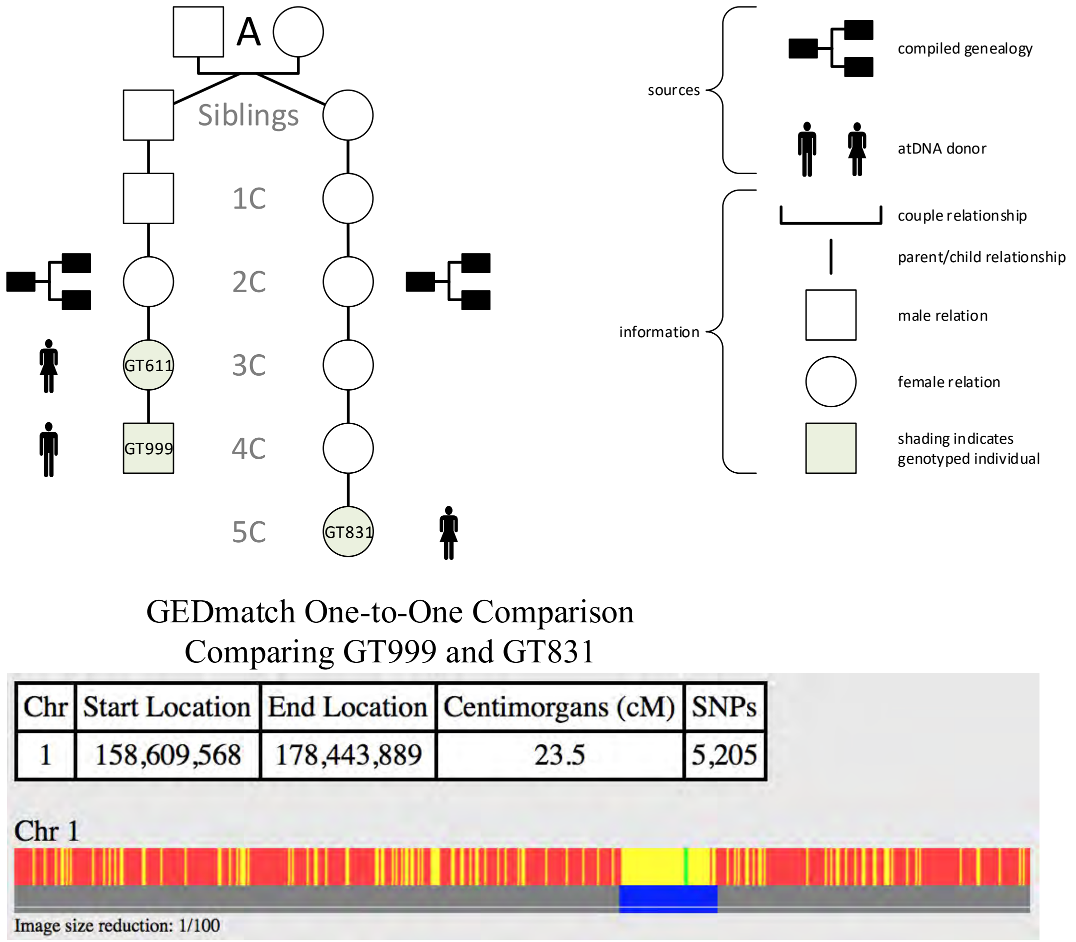
What’s Next?

Having isolated and correlated m and a that are true for the two *source* individuals, the genetic genealogist has the first piece of *evidence*—the first instance of $\mathcal{E}$—needed to answer the triangulation question.

A single instance of $\mathcal{E}$ is not sufficient for a conclusion. It is not sufficient to form a *hypothesis* that can be tested. This is only one piece of *evidence* that the matching segment in common was contributed by their common ancestor. At least one more instance of $\mathcal{E}$ is needed—an $\mathcal{E}'$ that correlates with $\mathcal{E}$—to have a *hypothesis* that can be tested. A second genotype matching the person of interest on the same segment and with a lineage to the same common ancestor is needed.

A Concrete Example

Consider the following example from research about GT999’s Chr1.



GEDmatch One-to-One Comparison
Comparing GT999 and GT831

Chr	Start Location	End Location	Centimorgans (cM)	SNPs
1	158,609,568	178,443,889	23.5	5,205

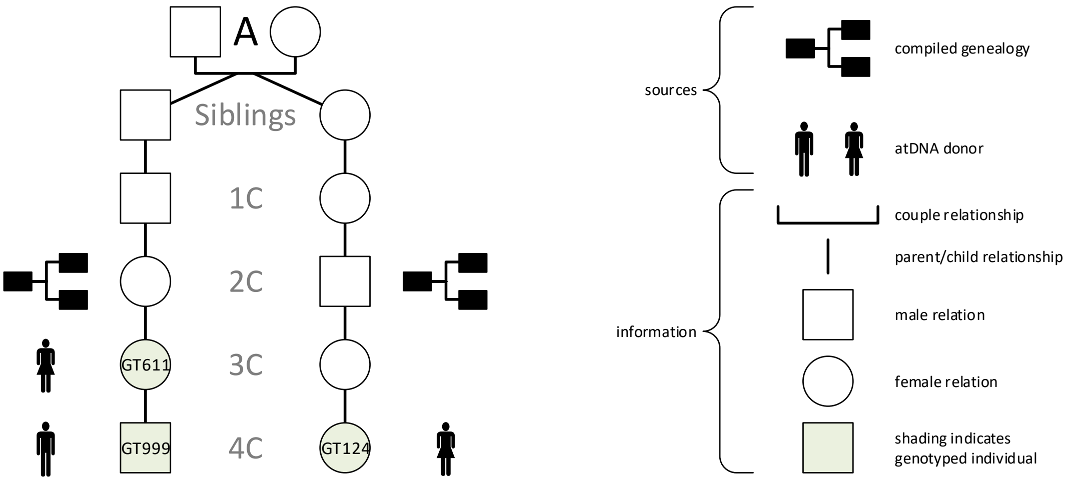
Chr 1

Image size reduction: 1/100

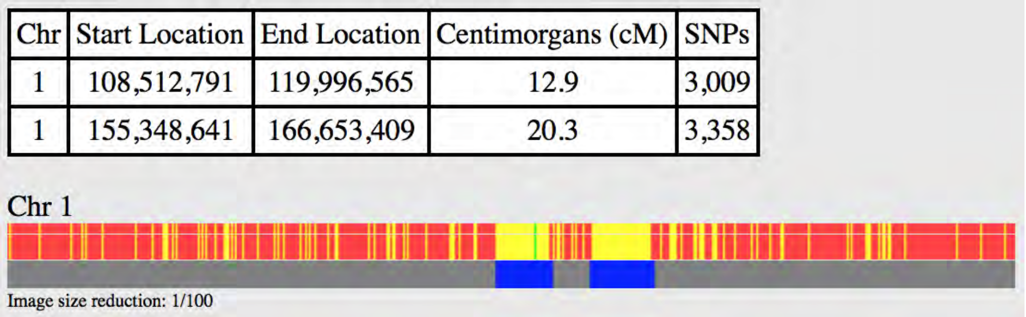
Figure 13: Graph representing the common ancestor between GT999 and GT831, and their GEDmatch comparison.[201,202,203]

In Figure 13, there is a person—identified as GT831—who matches GT999 on Chr1. The match is located between 159M and 178M, is 23.5 cM, and is represented by 5,205 SNPs. There is *information* to support supposition m. An examination of the compiled genealogies provided by GT999 and GT831 reveals that the couple—labeled *A*—contains the ancestor in common between GT999 and GT831. GT999 and GT831 are 4th cousins 1x removed. There is *information* to support supposition a. Because $m \wedge a \rightarrow \mathcal{E}$, one instance of $\mathcal{E}$—one piece of evidence—now exists to suggest that GT999 (and GT831) received the matching segment on Chr1 between 159M and 178M from one of their ancestors in the couple designated *A*.

But one item of *evidence* is not enough to have a valid *hypothesis*. A second instance of $\mathcal{E}$ is needed.



GEDmatch One-to-One Comparison
Comparing GT999 and GT124



Chr	Start Location	End Location	Centimorgans (cM)	SNPs
1	108,512,791	119,996,565	12.9	3,009
1	155,348,641	166,653,409	20.3	3,358

Chr 1

Image size reduction: 1/100

Figure 14: Graph representing the common ancestor between GT999 and GT124, and their GEDmatch comparison.[204,205,206]

In Figure 14, a person identified as GT124 matches GT999 on Chr1 in two locations. One of these locations—the segment located between 155M and 167M—overlaps with the location in common with GT831. If the segment of interest is changed to be the overlapping region shared between the three genotypes—the segment between 159M and 167M (14.0 cM)—there is *information* to support supposition m.[207] An examination of the compiled genealogies provided by GT999 and GT124 reveals that the same couple *A* is common between GT999 and GT124; they are 4th cousins. There is information to support supposition a. Again $m \land a \to \mathcal{E}$, and there is now a second instance of $\mathcal{E}$ suggesting that GT999 (and GT124) received the matching segment on Chr1 between 159M and 167M from an ancestor in the couple designated *A*.

CORRELATING THE INSTANCES OF $\mathcal{E}$

With two instances of $\mathcal{E}$, there is still one more condition that must be met to have a *hypothesis* that can be tested. Both instances of $\mathcal{E}$ must correlate with each other.

Correlating m

The correlation of m between the two instances of $\mathcal{E}$ is crucial. I cannot over-emphasize this point. Failing to do this correlation is the source of many errors and will render any conclusions invalid.

Genetic genealogists usually work with a group of genotypes that match on a single segment—herein called a triangulated group. This group is often identified prior to identifying a common ancestor. All the individuals in the triangulated group need to match the same segment (HIR).

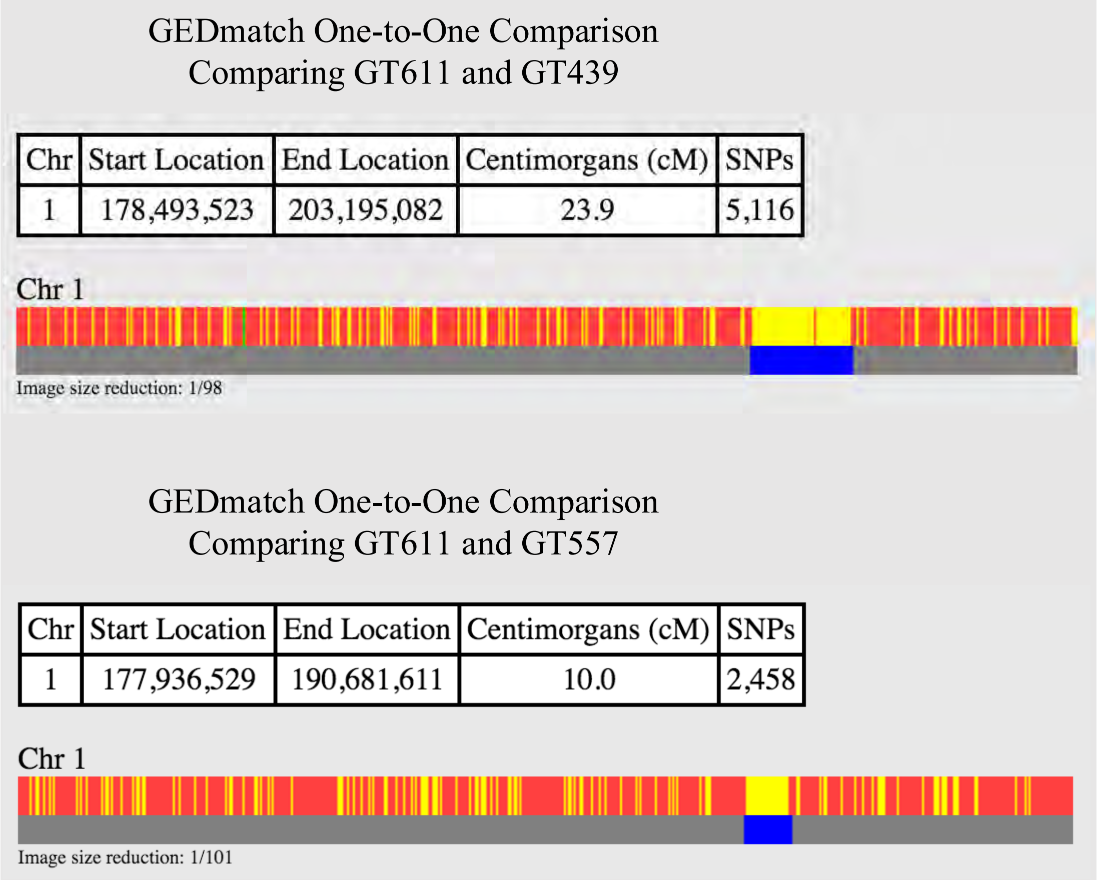
The question that needs to be answered is this: Is the atDNA in each piece of evidence from the same ancestor (even if we cannot identify which ancestor)? When considering a list of matches in a particular region, there are three possible matching scenarios. The segment could:

1. match a sequence of alleles along a person's paternal chromosome
2. match a sequence of alleles along a person's maternal chromosome
3. match a mix of maternal and paternal alleles in the region of interest, making it a false (IDC) match.

Without additional conclusions in place (like knowing one of the individuals is a paternal match), it is not possible to know which case is being encountered. However, *a third comparison—a correlation comparison—will indicate whether the three individuals match each other in the same way*.

If the genotype A matches genotype B and genotype C along a shared region of interest, a comparison that shows genotype B and genotype C matching each other over that same region ensures that all three individuals are matching the region of interest in the same way; it would also be a strong indication that the matching segment belongs to a parent haploid (case 1 or case 2 above). On the other hand, if genotype B and genotype C do not match each other, it cannot be true that the genotype A matches with genotype B and genotype C on the same chromosome. It must be that genotype B or genotype C each match separate parent chromosomes (one maternal and one paternal), or that one or both of the genotypes is falsely (IBC) matching genotype A.

As an example, consider GT611 who matches both GT557 and GT439 on Chr1 with an apparent overlap between 178M and 191M.



GEDmatch One-to-One Comparison
Comparing GT611 and GT439

Chr	Start Location	End Location	Centimorgans (cM)	SNPs
1	178,493,523	203,195,082	23.9	5,116

Chr 1

Image size reduction: 1/98

GEDmatch One-to-One Comparison
Comparing GT611 and GT557

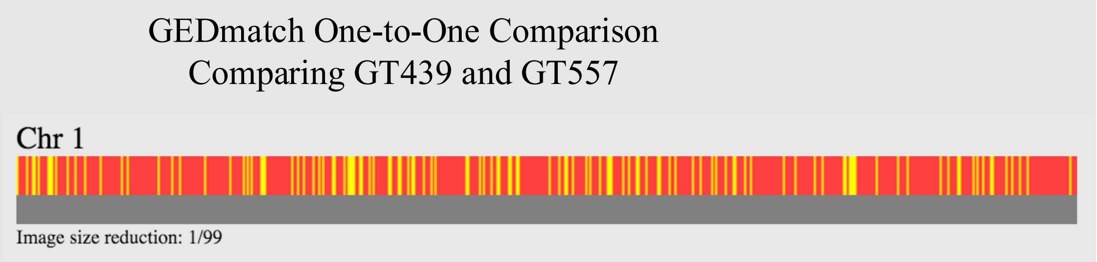
Chr	Start Location	End Location	Centimorgans (cM)	SNPs
1	177,936,529	190,681,611	10.0	2,458

Chr 1

Image size reduction: 1/101

Figure 15: Comparisons showing GT611's shared segments with GT439 and GT557 respectively.[208,209]

The comparisons in Figure 15 show the matches between GT611 and GT439 and GT557 respectively.



GEDmatch One-to-One Comparison
Comparing GT439 and GT557

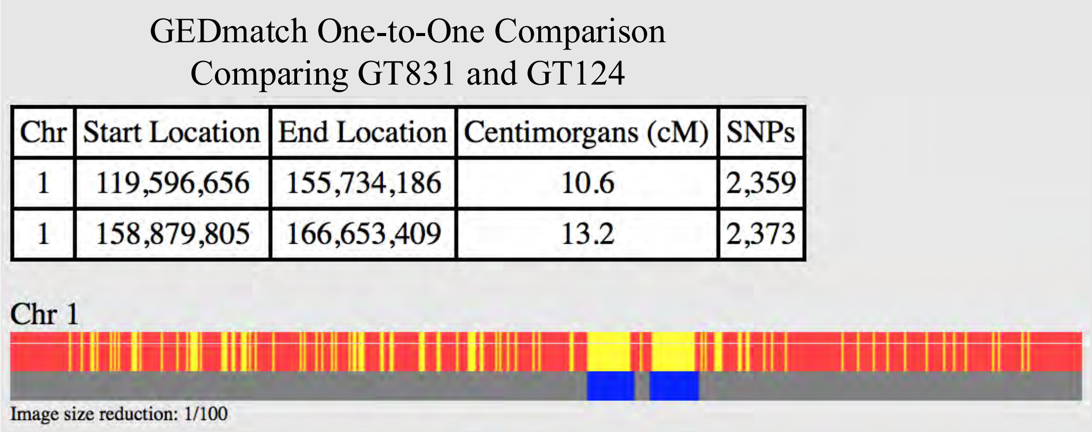
Chr 1

Image size reduction: 1/99

Figure 16: Correlation comparison between GT439 and GT557 showing no shared match.[210]

The correlation comparison—the comparison between GT439 and GT557—did not show a match (Figure 16). Thus, the instances of m do not correlate. [Research later showed that GT439 was a maternal match and GT557 was a paternal match for GT611.]

Copyright 2021. This work is licensed under a CC BY-NC-SA 4.0 license.



GEDmatch One-to-One Comparison
Comparing GT831 and GT124

Chr	Start Location	End Location	Centimorgans (cM)	SNPs
1	119,596,656	155,734,186	10.6	2,359
1	158,879,805	166,653,409	13.2	2,373

Chr 1

Image size reduction: 1/100

Figure 17: Correlation comparison between GT831 and GT124 showing that correlating match exists for the segment 159M to 167M.[211]

Returning to the case of GT999 matching GT831 and GT124 on Chr1 over the segment between 159M and 167M (see Figure 13 and Figure 14), the correlation comparison—comparing GT831 and GT124—reveals a match over the same segment (see Figure 17). The instances of m do correlate.

CORRELATING a

It is intuitive that both instances of ε correlate with regard to the common ancestor a. The same couple *A* was identified as the common ancestor for both instances.

TRIANGULATION IN TERMS OF ε

How is it that two or more instances of ε are the building blocks of triangulation?

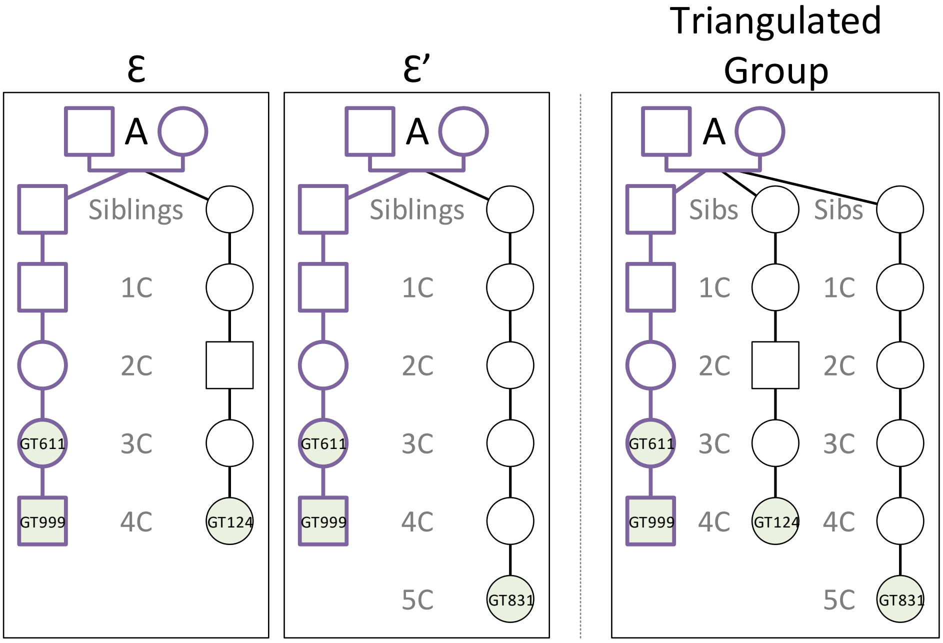


Figure 18: Two instances of ε are combined to form a triangulated group.

In the example of GT999 matching GT831 and GT124, two instances of *evidence* ($\mathcal{E}$ and $\mathcal{E}'$) have been identified. Note that the GT999's lineage is repeated in each representation of $\mathcal{E}$ (highlighted in lavender in Figure 18). If the representations of each instance of $\mathcal{E}$ are combined and the duplication removed, the result is a representation of a triangulated group, and the reason it is called triangulation begins to be apparent. *Appendix B* describes the testing of this triangulated group.

TESTING

With both instances of $\mathcal{m}$ correlating, and both instances of $\mathcal{a}$ correlating, it is safe to declare that both instances of $\mathcal{E}$ correlate. A *hypothesis* exists that can be tested.

For a *hypothesis* to become a *conclusion*, it must be scrutinized. No one *source* is foolproof. *Information* items from a *source* could be all right, all wrong, or a mix of the two (a common scenario in genealogical sources). It follows that an item of *evidence* based on *information* from a single *source* could be either right or wrong. Testing employs both analysis and correlation to evaluate the reliability of *sources*, *information*, and *evidence* used to form the *hypothesis* under scrutiny. Alone, analysis and correlation are insufficient; both types of testing must both be applied.

TESTS OF ANALYSIS

Tests of analysis examine the characteristics that affect the reliability of *sources*, items of *information*, and items of *evidence* in isolation. Tests will not prove correctness but will give insight into the likelihood of errors or misinterpretations. In genetic genealogy, tests of analysis may require an examination of biological possibilities, genome sequencing technologies and techniques, result reporting, feasibility to answer a particular question, donor motives, sample provenance, etc. Not all of these will be explored in this text. In typical scenarios, it is not necessary to examine such things as biological anomalies, provenance, and motives. The following tests of analysis should always be considered.

QUANTITATIVE CONSIDERATIONS

Given two genotypes, what can be said about the likelihood that sharing between them is IBD? What can be said about the likelihood that the sharing between them fits the relationship being examined? Some quantitative factors that affect the reliability of *conclusions* should be considered.

Matching Segment Size

Sharon and Brian Browning defined IBD in terms of shared haplotype frequency (i.e., rarity).[212] They point out that the amount of sharing (the size of the shared region) is generally used as a proxy measure of frequency—the principle being that the larger the shared region, the rarer it must be. The size of the region is important to one's confidence that the region is consistent with IBD-sharing.

Attempts to quantify the likelihood a matching segment is IBD use genotyped parent-child trios and measure the likelihood a segment shared with the child is also shared with at least one parent. As stated previously, a few have proffered anecdotal data, and John Walden has published findings from a larger dataset. A. J. Levin, using the published findings from Walden's dataset and regression, created the chart in Figure 19 to show the likelihood of IBD given matching segment sizes in cM.

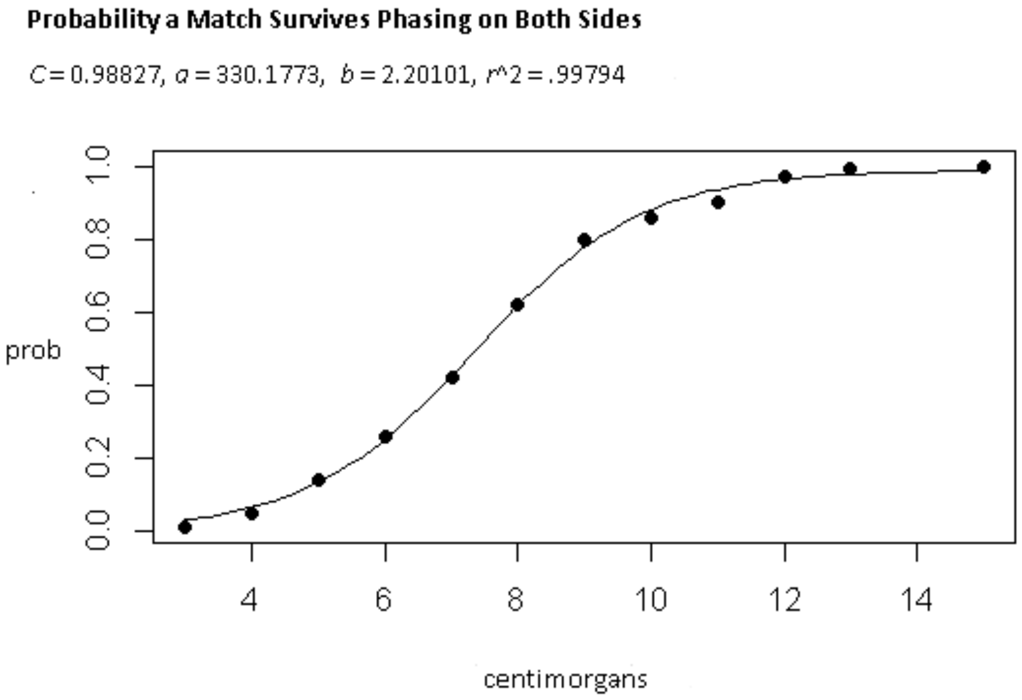


Figure 19: Probability a match survives when compared to a genotype phased with both maternal and paternal haplotypes.[213]

While Walden's findings have received criticism for lack of peer review, anecdotal evidence certainly fits well with the result. Figure 20 shows the number of matching segments, phased and un-phased, of a given size for GT999.

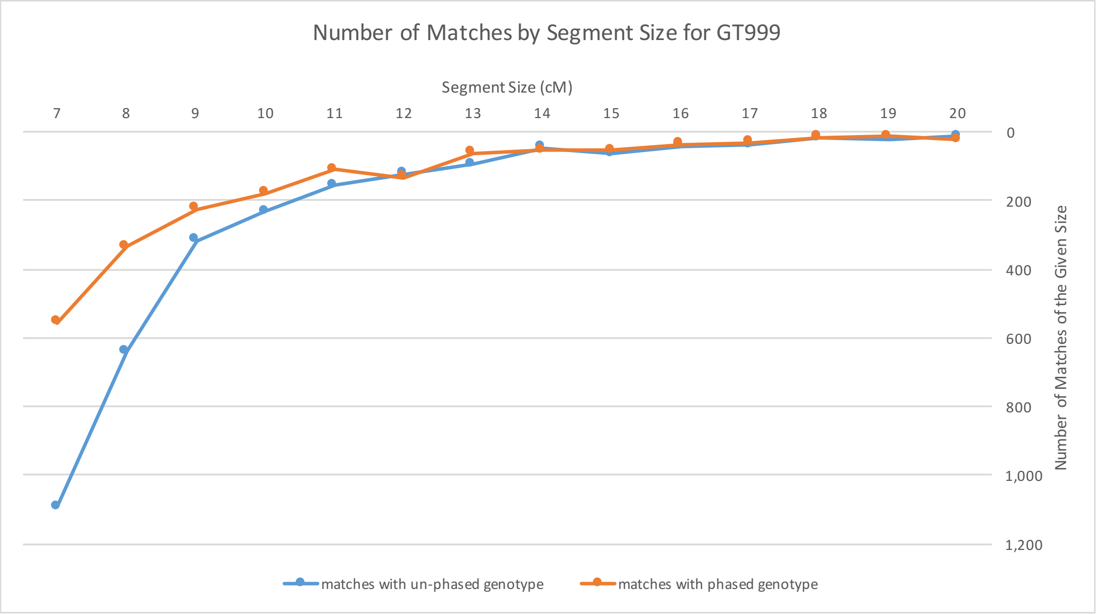


Figure 20: Number of matches by segment size for GT999.[214]

As match sizes grow larger, the number of segments of that size decreases—getting rarer. This aligns with the likelihood prediction of Figure 19. At the 7 cM size, nearly half of the segments that matched the un-phased genotype do not match the phased genotype. As match size increases, perhaps somewhere between 14-16 cM, the phased and un-phased genotypes produce roughly the same set of matches.

Even in cases where phased genotypes are used, current genotype data cannot be reliably used to detect IBD segments smaller than 5 cM. Durrand et al studied 2,952 parent-child-trios in a set of 25,432 genotyped individuals.[215] In their study, matching segments were declared IBD if a haplotype matching the child in a trio had at least an 80% segment overlap with at least one parent haplotype [perhaps a too-lenient standard?]. They state the following:

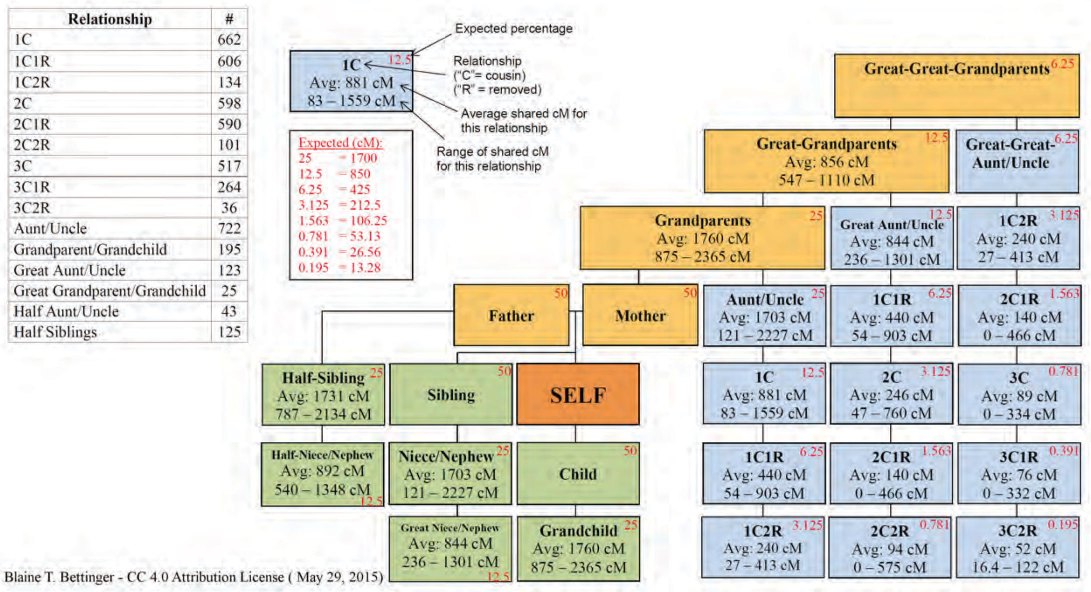
> "Most 2–3 cM segments are erroneous, and only segments longer than 5 cM have a negligible number of false positives. Indeed, when filtering solely by genetic length, all segments shorter than 5 cM must be discarded to achieve a precision value of 0.8."

As a quantitative consideration, matching segment size is important. If match size is not in a highly IBD-consistent range, conclusion reliability is poor. As Bettinger puts it: "…it is the responsibility of the genealogist to place a VERY high burden on any argument that utilizes … small segments….”[216]

Matching segment size is not just problematic for the genetic genealogist. Genotyping companies face a dilemma when selecting matching thresholds. Should the threshold be set high (matches reported out are mostly IBD-consistent; but many real matches are not reported—i.e., many false negatives), or set low (increases the number of real matches reported; but reports many non-cousins as potential matches—i.e., many false positives). Different research objectives might benefit from being able to tune matching segment size thresholds, but the reality is that genotyping companies control initial thresholds and genetic genealogists are left to operate with the data reported forward by these thresholds. Genotyping companies seem to target a threshold that balances the number of false-negatives and false-positives for matches being reported at the threshold size.[217]

Total Shared IBD

A second quantitative question asks: Is the size of the match a reasonable fit with the relationship being tested with the *hypothesis*? This question can be answered by comparing the measured amount of sharing to the expected amount of sharing.



Relationship	#
1C	662
1C1R	606
1C2R	134
2C	598
2C1R	590
2C2R	101
3C	517
3C1R	264
3C2R	36
Aunt/Uncle	722
Grandparent/Grandchild	195
Great Aunt/Uncle	123
Great Grandparent/Grandchild	25
Half Aunt/Uncle	43
Half Siblings	125

Expected (cM):
25 = 1700
12.5 = 850
6.25 = 425
3.125 = 212.5
1.563 = 106.25
0.781 = 53.13
0.391 = 26.56
0.195 = 13.28

Expected percentage
Relationship ("C"= cousin) ("R" = removed)
Average shared cM for this relationship
Range of shared cM for this relationship
1C 12.5 Avg: 881 cM 83 – 1559 cM
Great-Great-Grandparents 6.25
Great-Grandparents 12.5 Avg: 856 cM 547 – 1110 cM
Great-Great-Aunt/Uncle 6.25
Grandparents 25 Avg: 1760 cM 875 – 2365 cM
Great Aunt/Uncle 12.5 Avg: 844 cM 236 – 1301 cM
1C2R 3.125 Avg: 240 cM 27 – 413 cM
Father 50
Mother 50
Aunt/Uncle 25 Avg: 1703 cM 121 – 2227 cM
1C1R 6.25 Avg: 440 cM 54 – 903 cM
2C1R 1.563 Avg: 140 cM 0 – 466 cM
Half-Sibling 25 Avg: 1731 cM 787 – 2134 cM
Sibling 50
SELF
1C 12.5 Avg: 881 cM 83 – 1559 cM
2C 3.125 Avg: 246 cM 47 – 760 cM
3C 0.781 Avg: 89 cM 0 – 334 cM
Half-Niece/Nephew Avg: 892 cM 540 – 1348 cM 12.5
Niece/Nephew 25 Avg: 1703 cM 121 – 2227 cM
Child 50
1C1R 6.25 Avg: 440 cM 54 – 903 cM
2C1R 1.563 Avg: 140 cM 0 – 466 cM
3C1R 0.391 Avg: 76 cM 0 – 332 cM
Great Niece/Nephew Avg: 844 cM 236 – 1301 cM 12.5
Grandchild 25 Avg: 1760 cM 875 – 2365 cM
1C2R 3.125 Avg: 240 cM 27 – 413 cM
2C2R 0.781 Avg: 94 cM 0 – 575 cM
3C2R 0.195 Avg: 52 cM 16.4 – 122 cM

Blaine T. Bettinger - CC 4.0 Attribution License (May 29, 2015)

Figure 21: Data from Blaine Bettinger's Shared cM Project.[218]

As discussed in *Evidence from Quantitative Information*, the amount of sharing for close relationships should fall in an expected range. While these ranges are not exactly defined, the ranges from Bettinger’s *Shared cM Project* (Figure 10 and Figure 21) can provide practical guidance for these close-in relationships. In particular, the shapes of the distributions evident in Figure 10 will help exclude outliers that are reported in Figure 21. These outliers most likely represent data entry errors, or misidentified relationships.[219]

Tim Janzen continues to analyze expected amounts of sharing for relationships that are more distant—relationships involving 5th generation ancestors.[220] Janzen makes comparisons using 5 cM and 4 cM match thresholds and creates separate totals for matching segments that appear in these reports. If there are family groups involved (see Figure 26 and Figure 27 and the discussion related to independence), Janzen gets these same totals for each individual in the family group, then averages the totals for all of the individuals in the group (in the same generation) to create totals that represent the group. He then compares these totals with the expected amount. (See Figure 36 in *Appendix B* for an example.) Often his totals seem high, but this would be expected as some of the smaller matches included by the 5 cM and 4 cM thresholds will not be IBD. The *Shared cM Project* averages can be higher than expected for this same reason. Another reason that the averages can be higher than expected is due to additional background genealogical relationships that may not be known to the genetic genealogist. This can cause additional IBD segments to be included in the totals. Sometimes accurate chromosome mapping can identify these shared segments and they can then be excluded from the averages.

As research moves to common ancestors beyond the 5th generation, the expected amount of sharing is not so readily identified. The halving of the expected amount of shared atDNA would predict sharing at 3.32 cM for a fifth cousin—an amount that has already been flagged as untenable given current genotyping.[221] In contrast, GT999 appears to share 29.6 cM with his fifth cousin GT880.[222,223,224] Is this unreasonable because it more closely approximates the expected amount of sharing for a third cousin once removed? A researcher might find this match hard to accept because it seems incongruous to the expectations. Additionally, it is clear from the data in Figure 21 (and Figure 11) that some cousins in the third cousin range (and beyond) do not share any atDNA. This raises two questions. What is the probability that two people will share any atDNA IBD through a given ancestor? If sharing exists, how much sharing is expected?

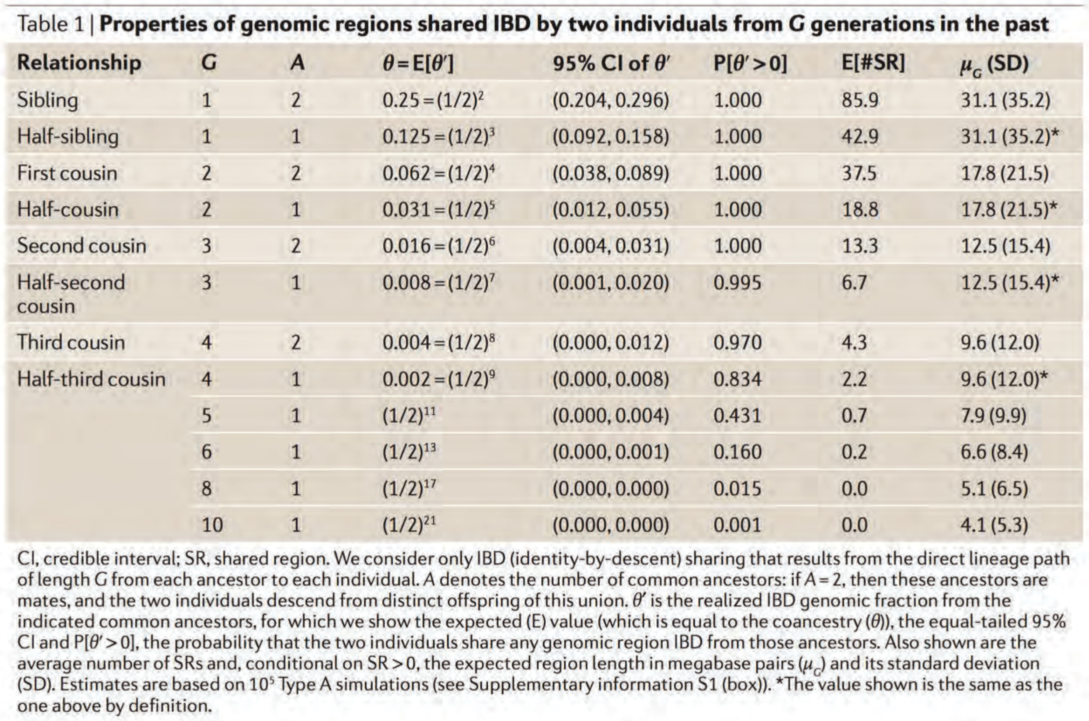


Table 1 | **Properties of genomic regions shared IBD by two individuals from G generations in the past**

Relationship	G	A	$\theta = E[\theta']$	95% CI of θ'	$P[\theta' > 0]$	E[#SR]	μ_G (SD)
Sibling	1	2	$0.25 = (1/2)^2$	(0.204, 0.296)	1.000	85.9	31.1 (35.2)
Half-sibling	1	1	$0.125 = (1/2)^3$	(0.092, 0.158)	1.000	42.9	31.1 (35.2)*
First cousin	2	2	$0.062 = (1/2)^4$	(0.038, 0.089)	1.000	37.5	17.8 (21.5)
Half-cousin	2	1	$0.031 = (1/2)^5$	(0.012, 0.055)	1.000	18.8	17.8 (21.5)*
Second cousin	3	2	$0.016 = (1/2)^6$	(0.004, 0.031)	1.000	13.3	12.5 (15.4)
Half-second cousin	3	1	$0.008 = (1/2)^7$	(0.001, 0.020)	0.995	6.7	12.5 (15.4)*
Third cousin	4	2	$0.004 = (1/2)^8$	(0.000, 0.012)	0.970	4.3	9.6 (12.0)
Half-third cousin	4	1	$0.002 = (1/2)^9$	(0.000, 0.008)	0.834	2.2	9.6 (12.0)*
	5	1	$(1/2)^{11}$	(0.000, 0.004)	0.431	0.7	7.9 (9.9)
	6	1	$(1/2)^{13}$	(0.000, 0.001)	0.160	0.2	6.6 (8.4)
	8	1	$(1/2)^{17}$	(0.000, 0.000)	0.015	0.0	5.1 (6.5)
	10	1	$(1/2)^{21}$	(0.000, 0.000)	0.001	0.0	4.1 (5.3)

CI, credible interval; SR, shared region. We consider only IBD (identity-by-descent) sharing that results from the direct lineage path of length G from each ancestor to each individual. A denotes the number of common ancestors: if A = 2, then these ancestors are mates, and the two individuals descend from distinct offspring of this union. θ' is the realized IBD genomic fraction from the indicated common ancestors, for which we show the expected (E) value (which is equal to the coancestry (θ)), the equal-tailed 95% CI and $P[\theta' > 0]$, the probability that the two individuals share any genomic region IBD from those ancestors. Also shown are the average number of SRs and, conditional on SR > 0, the expected region length in megabase pairs (μ_G) and its standard deviation (SD). Estimates are based on 10^5 Type A simulations (see Supplementary information S1 (box)). *The value shown is the same as the one above by definition.

Figure 22: Properties of genomic regions shared IBD by two individuals G generations in the past.[225]

Speed and Balding address these two questions. In Figure 22, column "P[θ' > 0]" gives the probability that two individuals will share any region IBD from the ancestor *G* generations in the past that is common between them. If a region is shared IBD, column "μ_G (SD)" gives the average size of that shared region with its standard deviation in parenthesis.

Be sure to note that all of the size information given by Speed and Balding is in Mbp.

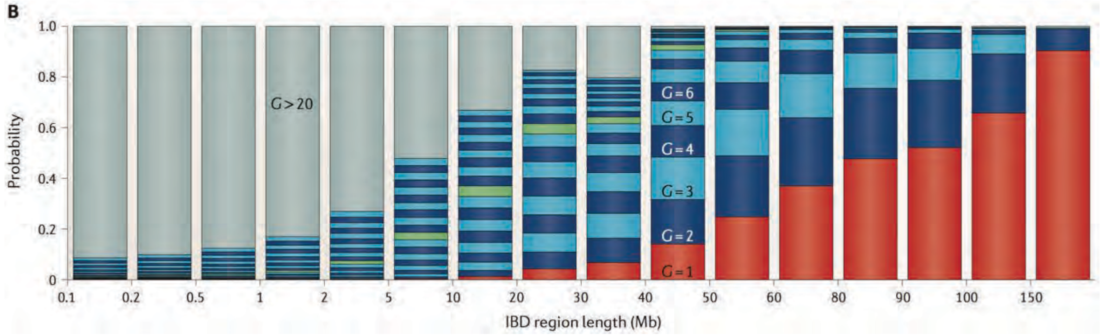


Figure 23: Statistics of IBD genomic regions.[226] *For all IBD regions arising from a common ancestor within the last 50 generations, the bars show how the distribution of the generation of the common ancestor depends on the length of the region. From bottom to top in the graph, the tranches correspond to G = 1 (red), G = 2...9 (alternating dark and light blue), G = 10 (green), G = 11...20 (alternating dark and light blue) and G >20 (grey).*

Figure 23 is a visual representation of Speed and Balding's findings regarding the possible common ancestors that might be associated with IBD segments of a given

© 2021. This work is licensed under a CC BY-NC-SA 4.0 license.

size. In the case of GT999 and GT880, the shared segment is 23,685,019 Mbp in length. Using the graph in Figure 23, there is about a 40% chance that the common ancestor associated with a shared segment of this size is more than 10 generations away from these two cousins, so it is possible for cousins at *G*=6 to share a segment of this size.

EVALUATING COMPILED GENEALOGIES

A compiled genealogy is typically an authored *source*. In some cases, however, genealogies are derivatives—copies of someone else's genealogy—and an in-depth analysis might require hunting down the original authored *source*. Questions that might be considered when examining a compiled genealogy include:

1. What motivated the creation of the compiled genealogy? Hobby? Society membership? Something else? Does this motivation make the research susceptible to bias?
2. Was the compiler careful and professional in their work?
3. Is the genealogy sourced? Were the best and most accurate sources used?
4. Was/is the genealogy open to challenge and correction?
5. How complete is the genealogy? Are there gaps? How many generations are complete?

The question of completeness is of particular importance in the hunt for the most recent common ancestor (MRCA).

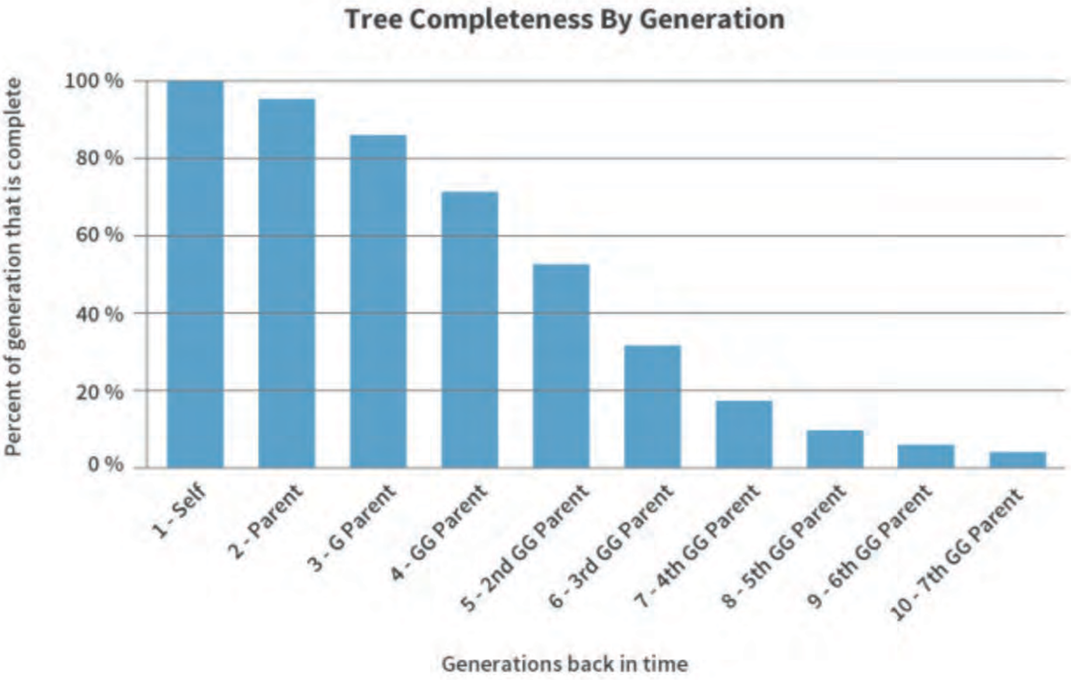
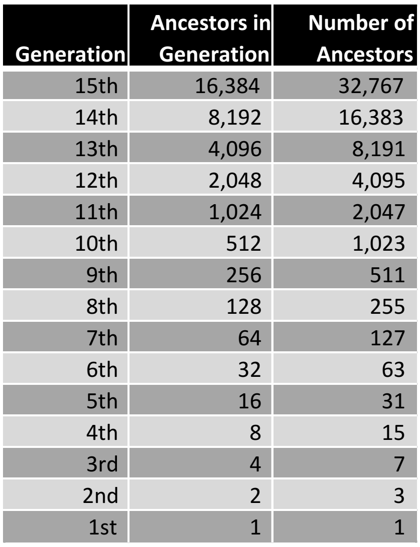


Figure 24: Sample of genealogies studied in AncestryDNA™ DNA Circles™ White Paper.[227] *Bars show the proportion of ancestors whom, for that generation of a pedigree, are known and documented (averaged over all studied pedigrees).*

Data about pedigree completeness provided by AncestryDNA™ (Figure 24) highlights this issue. The 512 seventh great-grandparents that need to be identified when seeking a MRCA in the 10th generation (Figure 25) is a hefty requirement and is, in many cases, impossible to satisfy. Being able to evaluate only 500 of the possible 1,023 pedigree slots makes a declaration of a single MRCA susceptible to doubt. *Conclusions* based on the *information* from such genealogies are made weaker by their incompleteness.



Generation	Ancestors in Generation	Number of Ancestors
15th	16,384	32,767
14th	8,192	16,383
13th	4,096	8,191
12th	2,048	4,095
11th	1,024	2,047
10th	512	1,023
9th	256	511
8th	128	255
7th	64	127
6th	32	63
5th	16	31
4th	8	15
3rd	4	7
2nd	2	3
1st	1	1

Figure 25: Number of ancestors (pedigree slots) in each generation, and the total number of ancestors that need to be searched (by generation) when seeking to identify a MRCA at that generation.

Copyright 2021. This work is licensed under a CC BY-NC-SA 4.0 license.

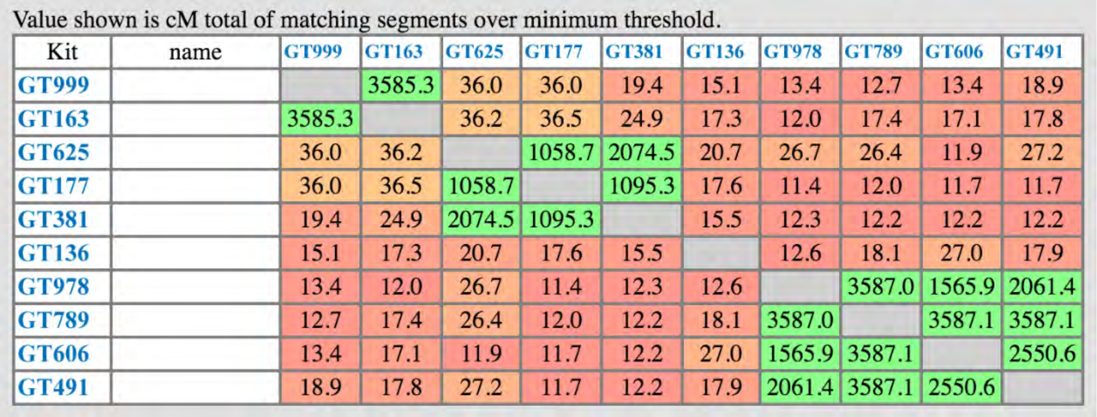
TESTS OF CORRELATION

Tests of correlation examine whether independent items are in agreement. Items in agreement may become *conclusions*. Items in disagreement are in conflict. Conflicts must be resolved before *conclusions* are possible.

INDEPENDENCE VS. RELATEDNESS

Correlation is only valid if the items being compared are independent. Independence connotes separate and distinct informants. Related items come from the same informant. It does not make sense to correlate related items; these items are just alternate representations of the original, and the original will always correlate with itself.

Independence is a key consideration when triangulating. The lineages to the common ancestor need to be independent. Having three (or more) independent lineages to the common ancestor are what gives triangulation its power to prove relationships.



Value shown is cM total of matching segments over minimum threshold.

Kit	name	GT999	GT163	GT625	GT177	GT381	GT136	GT978	GT789	GT606	GT491
GT999			3585.3	36.0	36.0	19.4	15.1	13.4	12.7	13.4	18.9
GT163		3585.3		36.2	36.5	24.9	17.3	12.0	17.4	17.1	17.8
GT625		36.0	36.2		1058.7	2074.5	20.7	26.7	26.4	11.9	27.2
GT177		36.0	36.5	1058.7		1095.3	17.6	11.4	12.0	11.7	11.7
GT381		19.4	24.9	2074.5	1095.3		15.5	12.3	12.2	12.2	12.2
GT136		15.1	17.3	20.7	17.6	15.5		12.6	18.1	27.0	17.9
GT978		13.4	12.0	26.7	11.4	12.3	12.6		3587.0	1565.9	2061.4
GT789		12.7	17.4	26.4	12.0	12.2	18.1	3587.0		3587.1	3587.1
GT606		13.4	17.1	11.9	11.7	12.2	27.0	1565.9	3587.1		2550.6
GT491		18.9	17.8	27.2	11.7	12.2	17.9	2061.4	3587.1	2550.6	

Figure 26: Total amount of atDNA sharing between members of the group that triangulates on a matching Chr4 segment from 187M to 191M.[228]

Figure 26 shows the comparison of ten individuals that match a segment on Chr4 (187M to 191M). The question is: How many independent answers to the triangulation question—instances of $\mathcal{E}$—are possible from this group? One might be tempted to say nine—one for each individual that can be paired with the researcher's person-of-interest. However, only three independent answers are possible because there are family groups (identifiable in Figure 26 as rows with green clusters on the diagonal) that contribute related information and cannot be considered independently.

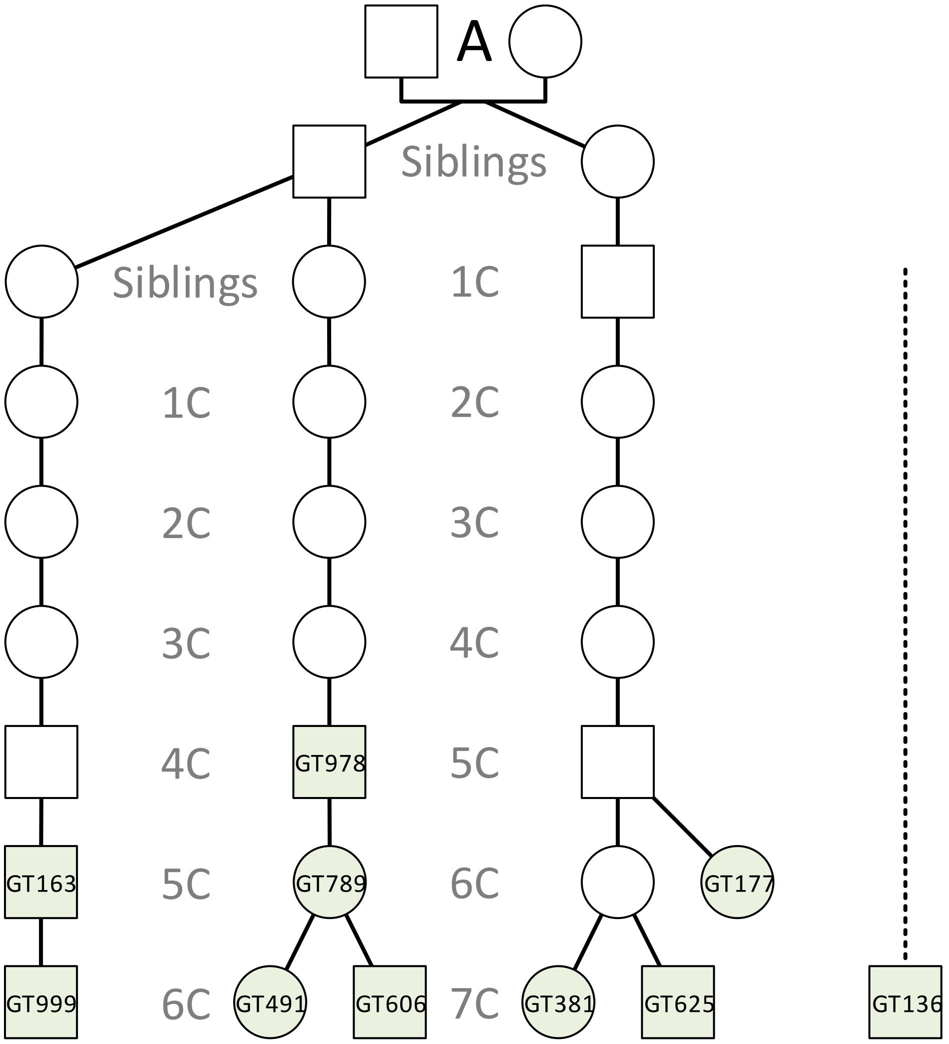


Figure 27: Representation of the lineages (known and unknown) in a triangulated group for a Chr4 segment (187M to 191M).[229,230,231]

The first group (GT163 and GT999) are a father/son pair (see Figure 27). The son received his copy of the matching segment from his father. The son cannot add any independent information about the lineage between the father and the common ancestor because he shares the same lineage (and the same set of meiosis events). The son's answer is derived from the father's answer and is, therefore, related.

The second group (GT978, GT789, GT491 and GT606) is a father, his daughter, and two of her children (see Figure 27). The mother and her children in this family group received their copy of the matching segment from GT978. The answers that GT789, GT606 and GT491 would provide would all be related to the answer that GT978 provides.

The third group (GT177, GT625 and GT381) received their copy of the matching segment from their father/grandfather (see Figure 27). For the subsequent six generations that lead to the common ancestor from their father/grandfather, these three can add no independent information; their answers are identical—from the same source.

The only other possibly independent answer could come from GT136 who has no close familial relationships with other members of the group. His exact relationship to the group remains unknown.

Even though this triangulated group has ten people that match the segment being considered, only three independent instances of $\mathcal{E}$ are possible. *Appendix C* describes the testing for this triangulated group.

TESTING THE IBD ASSERTION

One of the primary assertions that needs to be accepted is that the matching segment being considered was received IBD. Tests of correlation are crucial in establishing such claims. If a match stands up to quantitative testing (size is consistent with IBD, shared total is plausible), there are a number of ways that correlation is used to show/deny plausibility that a matching segment was received IBD.

PHASED MATCHING

One of the best mechanisms for eliminating IBC matches is to make comparisons with phased chromosome data. For example, GT999 has 7933 matches over the default GEDmatch thresholds (matching segments of at least 7.0 cM and made up of at least 700 SNPs); using his phased genotypes (created with genotypes from both parents) reduced the number of matches to 2660.[232] Two thirds of the default matches were eliminated as IBC matches.

AncestryDNA™ uses phasing to help with their matching algorithm.[233] GEDmatch allows the users to generate and use phased data.[234] Tim Janzen, David Pike and Felix Immanuel have published utilities to phase genotypes.[235]

The best phased genotypes are based on child-parent trios; phasing is still possible with a child-parent duo. It is possible to phase data without a parent; Janzen has published information about this process.[236]

Using phased data ensures that the allele sequence used in comparisons is an allele sequence that actually exists on a single haploid chromosome received from a parent, not just a random mash-up of alleles from both parental chromosomes.

Generational Matching

If genotype has parents and/or grandparents in line with the lineage of the proposed common ancestor that are also genotyped, comparing the match with each related genotype in line with the match can be telling.

In Figure 27, consider the family group made up of GT978, GT789, GT491 and GT606. The fact that GT491 and GT606 can show they received the matching segment IBD from their mother, and that GT789 can show she received it IBD from her father gives credence to the claim that GT978 received it IBD from his ancestors.

If, instead, the grandfather and a grandchild both share the match, but the mother does not, how could it be that the matching segment was received IBD? This is a biological impossibility. The grandchild could not have received it IBD unless the mother also shares the match.

Intermediate Common Ancestors

The concept of an intermediate common ancestor is related to generational matching. Jim Bartlett wrote a notable blog on this topic.[237] In Figure 28, consider a hypothesis that GT654 received the identical segment under consideration IBD.

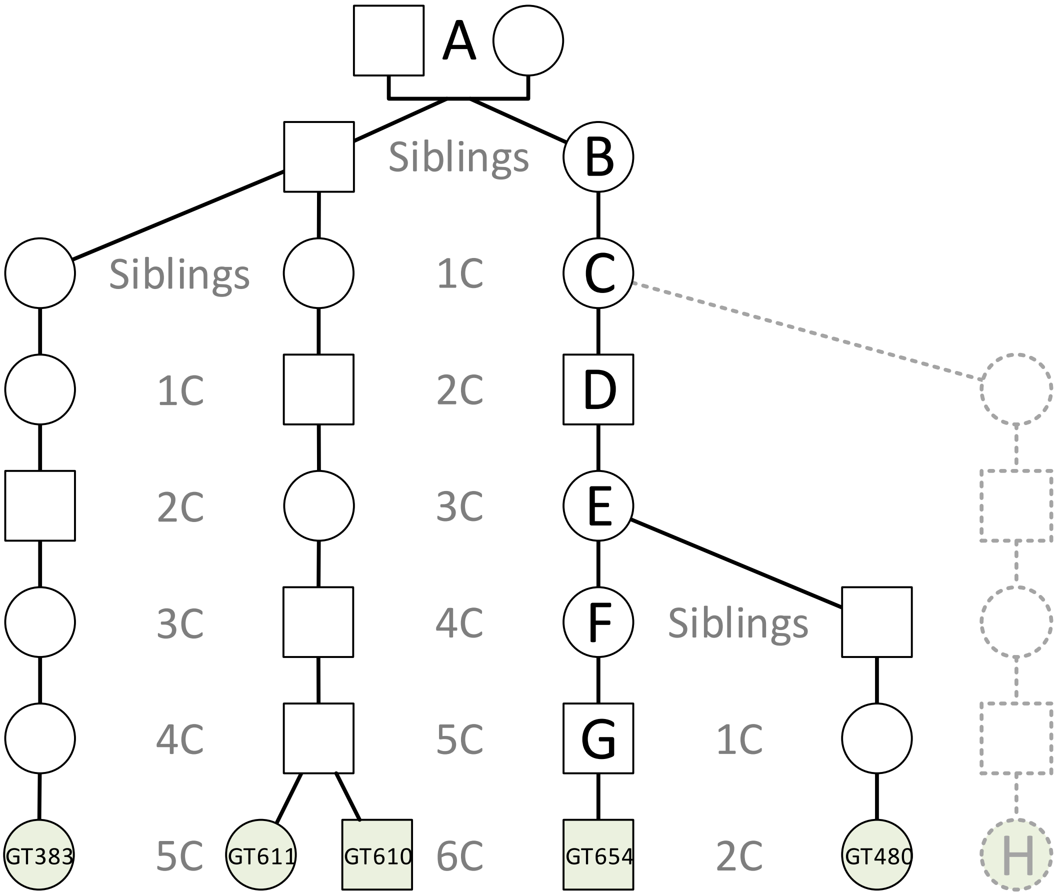


Figure 28: Representation of the lineages in a triangulated group for a Chr1 segment (177M to 191M).[238,239,240,241,242,243] *See also* Appendix D.

GT654 has a second cousin GT480 that shares the match in consideration. Considered in isolation, this match and the associated lineage that relates GT654 to GT480 qualifies as an instance of $\mathcal{E}$. This intermediate instance of $\mathcal{E}$ gives *evidence* that GT654 and GT480 received this matching segment IBD from their great-grandfather (labeled *E*). What if another independent cousin (*H*) existed that shared this match with a lineage that related GT654 and *H* via their ancestor *C*—an instance of $\mathcal{E}$ giving *evidence* that GT654 and *H* received this matching segment IBD? These intermediate instances of $\mathcal{E}$ would give considerable strength to the supposition that GT654 received the matching segment IBD from the triangulated group's common ancestor *A*. What if additional instances of $\mathcal{E}$ existed for B, D, F and G? It would be like knowing the provenance of that matching segment through every generation from *A* to GT654. Jim Bartlett likened this concept to "[walking] the segment back" through the generations to the common ancestor.[244]

Does a lack of intermediate common ancestor matches invalidate the *hypothesis* being tested? No. But they are desirable. Without them, there may be more risk in declaring a *conclusion*, but no more risk than declaring a *conclusion* based on the *evidence* that initiated testing.

Can intermediate common ancestors be sought out? Yes…and no. Yes, because it is sometimes possible to identify that a match may be in common with the person of interest because they may have a particular intermediate common ancestor in common with the person of interest. Given a list of all your n^th cousins, the probability of picking the cousin(s) that share the matching segment is small; but if the person of interest has a large number n^th cousins contributing genotypes, the probability of finding a match increases.[245] In other words, the probability of finding one by picking one is small, whereas the probability of finding one by comparing one's genotype to a whole database of genotypes is much greater.

Close Relative Matching

Close relatives that are related to the common ancestor—siblings, aunts, uncles, close cousins, etc.—can give *evidence* that a matching segment was received IBD.

In Figure 27, consider the family group identified by GT177, GT625 and GT381. The fact that GT381 has a sibling and an aunt that share the match in this group is *evidence* supporting the claim that she received her copy of the matching segment by descent from her more distant ancestors.

Conversely, GT136 has four relations that descend from his paternal grandfather—GT357, GT787, GT502 and GT241— that do not share the match being considered in the Figure 27 triangulated group.[246] Additionally, searching for a common ancestor among his paternal grandfather's ancestors has not produced any ancestors in common with the other members of the group. This becomes *evidence* that GT136 did not receive this matching segment IBD from his paternal grandfather.

Testing by matching with close relatives is just a special case of the *intermediate common ancestor* test. It is, perhaps, not natural to consider them as such because, often, no thought is given to "proving" the genetic relationship because it is "known" as such.

Match Stability

When considering generational matches, the stability of the match needs to be considered. Stability has to do with whether the matches conform to biological principles as it passes from generation to generation.

Consider the match shared with GT208 as it passes through four generations—from GT118 to her son GT163 to his son GT999 to his son GT186:

Kit #	Chr	Start Location	End Location	cM	SNPs
GT118 with GT208	1	63,175,608	76,807,359	12.1	3,132
GT163 with GT208	1	63,496,685	74,958,148	10.0	2,553
GT999 with GT208	1	63,336,657	74,958,148	10.2	2,609
GT186 with GT208	1	63,592,864	77,595,616	12.4	3,133

Figure 29: Unstable generational match with GT208.[247]

If the matching segment is IBD, biology would dictate that GT118 has the longest matching segment with GT208, and that successive generations are the same or smaller. Yet, the longest matching segment is with the most recent generation, and match size increases as it is transmitted from the 2nd generation to the 3rd and 4th generations—biological impossibilities. This match is apparently not stable.

Alternatively, it could be that the matching segment considered in Figure 29 is partially IBD and partially IBC. Perhaps the region between positions 63,592,864 and 74,958,148 is mostly (or completely) IBD, while the regions between positions 63,175,608 and 63,592,864 and between positions 74,958,148 and 77,595,616 are IBC. This may, in fact, be a more likely explanation for the example given in Figure 29. Perhaps additional questions could be asked. Perhaps other tests will assuage (or accentuate) the concerns.

Now consider the same generational matching with GT293:

Kit #	Chr	Start Location	End Location	cM	SNPs
GT118 with GT293	12	90,999,825	100,870,206	12.8	2,690
GT163 with GT293	12	90,999,825	100,870,206	12.8	2,691
GT999 with GT293	12	90,999,825	100,870,206	12.8	2,690
GT186 with GT293	12	90,999,825	100,870,206	12.8	2,668

Figure 30: Stable generational matching with GT293.[248]

With GT293, the matching segments are almost identical across all four generations. The only differences in the comparisons are in the number of identical SNPs in the matching sequence. This matching segment seems likely to be mostly or entirely IBD.

Chromosome Map Correlation

A chromosome map associates matching segments of atDNA with specific ancestors. If the chromosome in consideration has been mapped for the person of interest, the matching segment in consideration must fit within the map without conflict. The matching segment cannot inappropriately span known crossover locations. The common ancestor must fit as a relation of the known ancestors already associated with that location on the chromosome. If these conditions are not true, there is a conflict that must be resolved before a *conclusion* can be declared.

No one mapping technique will fully populate a given chromosome map. This means that chromosome maps are created over time and are a compilation of many *conclusions* accumulated over time. It follows that not every *hypothesis* can be tested against a map. But every *conclusion* could be added to a map and used in future evaluations.

In cases where the genotypes of three or more siblings are available, it is possible to jumpstart the creation of a chromosome map that shows matching segments (HIRs) received from their grandparents without accumulating a set of triangulated *conclusions* to do so. Kathy Johnston describes this method.[249] The process involves comparing the sibling sequence data to identify crossover locations, assigning crossovers to specific siblings, and then using logical inferences regarding FIRs, HIRs, and regions with no matching to work out HIRs received from each grandparent for each sibling. Figure 31 shows the Chr1 comparisons for four siblings and the maps that resulted. Note there are still portions of the map which could not be assigned. Also note that without knowing at least two *conclusions* (one about a paternal ancestor, and one about a maternal ancestor) that fit into the map, it is impossible to know which color belongs to which grandparent.

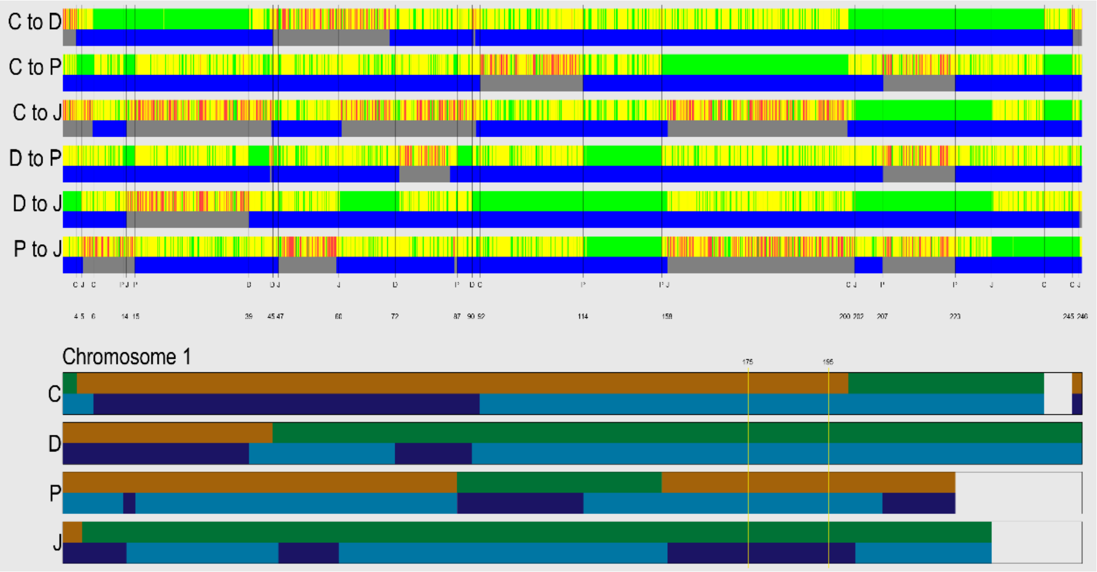


Figure 31: Chromosome 1 comparisons among four siblings (six bands at the top) and resulting maps showing half-identical regions inherited from grandparents (four maps at the bottom).[250,251,252]

The segment shared by the triangulated group represented in Figure 28 represents a paternal ancestor of GT611. Her Chr1 map is labeled *C* in Figure 31. It turns out her brother GT610 (labeled *P*) also matches this segment. With this information, it can be inferred that the top lines in these maps represent the chromosomes inherited from their father, and that the brown bands represent their paternal grandfather and the green bands their paternal grandmother.

Now consider the shared segment from Figure 28 in the context of the maps. The segment's position falls between the yellow lines (overlaid onto the maps). As such, the segment does not straddle any crossover locations. Only two of the siblings matched the segment, and the map predicted the second match (meaning that GT610 had not been compared to the triangulated group until after the map had been created and the maternal and paternal chromosomes had already been identified). The segment shared in Figure 28 fits without conflict into GT611's Chr1 grandparent map.

Excessive Matching

Some regions of the genome tend to match other genomes at a higher-than-expected rate—they are prone to excessive matching.[253] These regions have lots of identical matches. These matches are not identical because of recent shared ancestry but are considered the result of demographic or historic factors. In the genetic genealogy community vernacular, these groupings of excessive matches are termed *pile-ups*. The “Excess IBD sharing” section of the *Identical by descent* article on the ISOGG wiki gives several references to research about specific pile-up regions.[254]

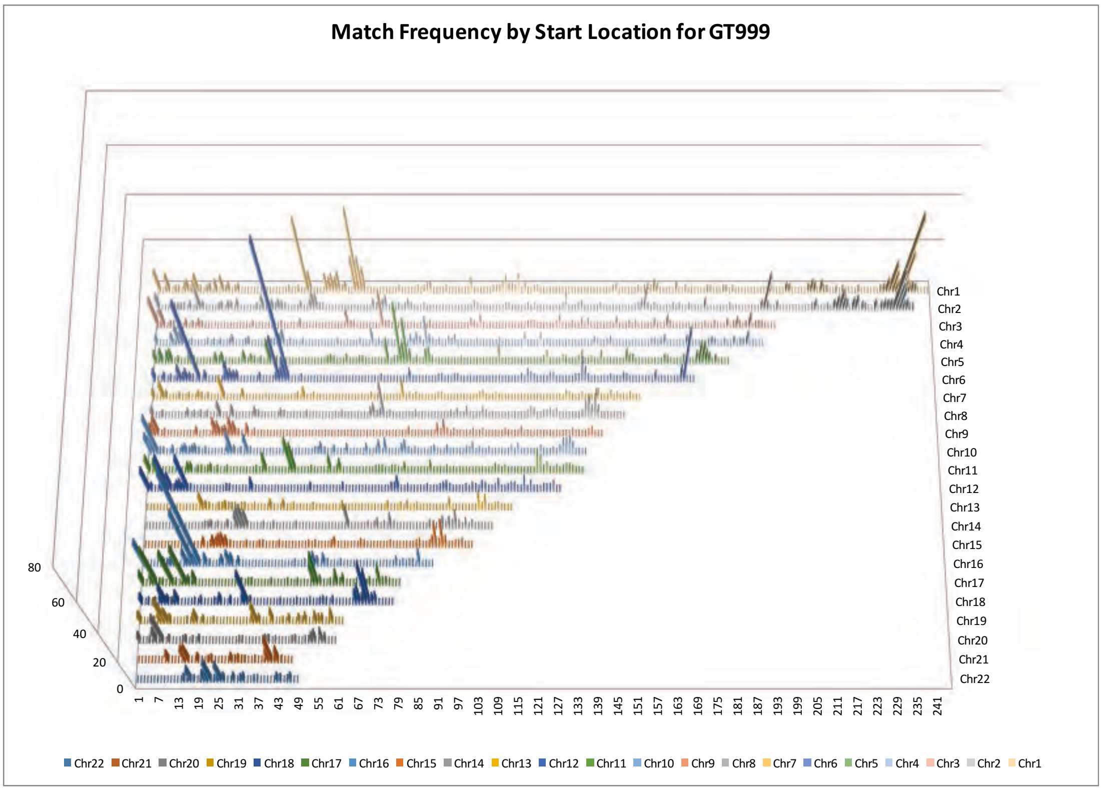


Figure 32: Plot of match frequency (1 Mbp tranches) to show potential pile-up regions for GT999.[255]

Figure 32 plots GT999’s match frequency by chromosome and start location. There are several locations with potential pile-ups. Chr1, Chr2, Chr5, Chr6 and Chr16 (and others) all have small regions with significantly higher numbers of matches compared to the rest of the genotype, made evident by tall spikes in the number of matches at locations along those chromosomes.

If one finds there are a lot of matches for a particular location in the genome, can it be assumed to be excessively matching? There could be reasons to answer no. It may be that there are several family groups (clusters of related genotypes) that make the frequency higher than expected. Sometimes these clusters include duplicate genotypes. Each cluster needs to be evaluated separately.

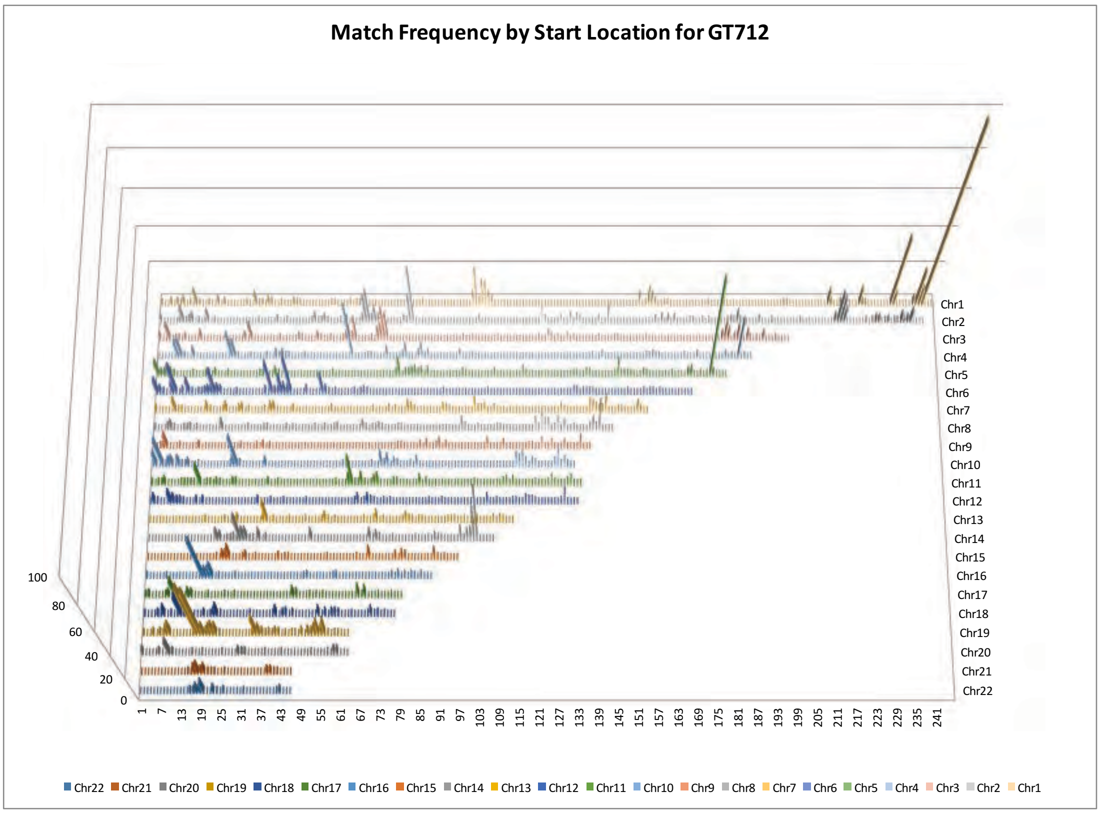


Figure 33: Plot of match frequency (1 Mbp tranches) to show potential pile-up regions for GT712.[256]

These pile-up regions are not necessarily consistent from genome to genome (note both the similarities and differences between Figure 32 and Figure 33), nor is there perfect agreement about what should be done about them. For example, AncestryDNA™ has added the so-called "Timber" algorithm to its processing to de-emphasize regions with excessive matching.[257] Some angst has been expressed about this addition, with many asserting that the algorithm eliminates useful matches.[258] Yet current limitations—both in genotyping and in the timeframes that genealogical records can viably cover—seem to prevent correlation or inference involving these matches. Dan Edwards asserts that incorporating these matches into a genealogical research process is futile—genealogically intractable.[259]

If a region seems to be excessively matched, one needs to consider whether such matches can even be considered IBD. Sharon and Brian Browning define IBD in terms of shared haplotype frequency.[260] If sharing exists between two haplotypes but is not likely with any other haplotypes, it is more likely IBD. The frequency that other haplotypes match a given region can indicate a likelihood that the region was received IBD. Excessive (i.e., high frequency) matching dilutes the supposition that region was received IBD.

Given a *hypothesis*, how many independent instances of $\mathcal{E}$ ought to be present in the *hypothesis*? Ideally, each match capable of giving an independent answer to the triangulation question should have an instance of $\mathcal{E}$ in the *hypothesis*. If the *hypothesis* should have four instances of $\mathcal{E}$ and only has three, the *hypothesis* likely remains tenable. If the *hypothesis* should have sixty-five instances of $\mathcal{E}$ and only has three, there is very likely reason for caution. If a *conclusion* is declared in this latter case, is there a basis to expect that further instances of $\mathcal{E}$ will not alter that *conclusion*? It seems likely to be an untenable position; the author's own foray into researching such a match (e.g., a triangulated group needing 25 instances of $\mathcal{E}$) ran afoul almost immediately with multiple, conflicting common ancestors.[261,262,263,264]

SOLUTION PREDICTS RELATIONSHIPS

Bartlett claims that match databases have doubled every fourteen months.[265] This means that new matches are continuously being added to match lists. From time to time, new matches will associate with triangulated groups that have previously *concluded* common ancestors. These matches are an opportunity to add additional instances of $\mathcal{E}$ in support of the existing *conclusion*. If the existing *conclusion* is sound, one would expect that the existing *conclusion* would predict the ancestor (or lineage) that would be found in common with the new match.

Finding the predicted ancestor (or lineage) in the compiled genealogy of the new match does not mean that the new instance of $\mathcal{E}$ does not need testing; all applicable tests of analysis and correlation should still be applied. If the predicted ancestor (or lineage) is still part of the resulting *conclusion* when testing is complete, it also stands to reason that the pre-existing *conclusion* was successful in predicting the new match's relationship to the triangulated group. This success is additional evidence that the matching segment being considered was received IBD.

COMMON ANCESTOR UNIQUENESS

Concluding that a particular matching segment of atDNA was received IBD from an ancestor is difficult if there is more than one viable candidate ancestor. Ideally, the compiled genealogies being searched would be completely known to the generation of the candidate ancestor, and the only ancestor common to these genealogies would be the candidate ancestor. The practicalities of reality merit consideration.

It is not uncommon for two pedigrees to have more than one candidate in common. If this is the case, additional instances of $\mathcal{E}$ are needed to narrow the field. Generally, the more independent instances of $\mathcal{E}$ that are present in the hypothesis, the more likely there will be only one candidate in common.

Sometimes, more instances of $\mathcal{E}$ result in conflicting candidates for a common ancestor. Is the common ancestor from an older generation? Perhaps there is an issue in the genealogies themselves—e.g., the genealogies are not complete, the genealogies are not accurate, or an unidentified NPE exists in the genealogies. Additional matches may be able to help sort out such conflicts.

Another reason for conflict is that the matching segment is not IBD. Should the match be categorized as excessively matching—IBS due to demographic or historical reasons? Is the match IBC?

It is common to encounter incomplete genealogies when searching for common ancestors. Often, a common ancestor cannot be identified in such cases—an incomplete instance of $\mathcal{E}$ (an $\mathcal{m}$ without an $\mathcal{a}$). Sometimes, like in the case of an adoptee, parentage is unknown, and a complete genealogy cannot exist until the parentage problem is solved. In some cases, the individual may withhold such information. In cases where the information is withheld, it may be possible to compile a genealogy as a surrogate of the missing information—though the privacy and protection of this information must be considered, and it may raise ethical concerns. An incomplete genealogy is probably not a reason to "fail" this test. It does marginalize the match's contribution to the hypothesis.

It is possible that a common ancestor will emerge even though one or more of the genealogies are incomplete. Incomplete genealogies mean incomplete data is used to determine the MRCA—leaving any identified MRCA open to question. It should be very rare, however, to have two separate instances of $\mathcal{E}$ (each with independent instances of $\mathcal{m}$ and independent lineages to $\mathcal{a}$). Therefore, having a number of independent instances of $\mathcal{E}$ in the hypothesis reduces the likelihood that an incomplete tree will result in an error due to incomplete genealogy.

LOOKING FORWARD

In considering hypothesis testing, this text has considered a number of the most common *tests of analysis* and *tests of correlation* to be used in validating and characterizing *conclusions* supported by atDNA *evidence*. Is this list exhaustive? No. Will additional tests be identified? Absolutely. This is an emerging practice in the genealogical community. Genetics and genomics are evolving fields of study. Further refinement of the tests and heuristics detailed herein are needed and expected; for example, it would be helpful to re-present the information reported by Speed and Balding in terms of genetic distance rather than physical distance so that it is more readily interpreted within the context that genetic genealogists use. Next-generation sequencing is sure to enable further innovation. This text presents a framework and methodology for identifying, evaluating, and presenting the strengths and weaknesses of a *conclusion* involving atDNA *evidence*.

CONCLUSION ACCEPTING

The desired outcome of *hypothesis* testing is a *conclusion*. If the *hypothesis* holds up to scrutiny—if it passes testing and all conflicts can be resolved—it becomes a *conclusion*. Testing safeguards the researcher from erroneous *conclusions*, and spotlights the strengths and/or weaknesses of the resulting *conclusion*.

For some tests, a failure will invalidate the *hypothesis*. For example, a *hypothesis* that does not pass a generational matching test cannot be accepted as a *conclusion*. Other tests only cast/remove doubt. An incomplete genealogy introduces doubt but does not invalidate. Doubt can be alleviated on the strength of other testing. If doubts mount, a *conclusion* may become untenable.

Testing may spotlight conflicts. Conflicts must be resolved or no *conclusion* can be made. For example, if a matching segment inappropriately spans a crossover location in a chromosome map, a *conclusion* is premature. It may be that the map itself is flawed and that fixing the map will resolve the conflict; if so, a *conclusion* may be possible; if not, the *hypothesis* must be discarded.

Extraordinary *conclusions* require extraordinary *evidence*.[266] In other words, the burden of proof is very high for *conclusions* that push the edges of possibility—*conclusions* that are unusual and/or improbable. At the same time, *conclusions* should not be discarded because all tests could not be applied, or because doubt remains. As much as possible, tests should be applied, and the outcomes discussed and weighed. As an exhaustive search and only two items of evidence in traditional genealogy can result in a conclusion, it follows that two scrutinized instances of $\mathcal{E}$ with no conflicts and no invalidating tests is sufficient to *conclude*—if explained.

Appendix F details two *conclusions* based on the triangulation of a segment shared by GT999 on Chr1 with two distant cousins. The author *concludes* that members of the triangulated group received the specified matching segment IBD from a particular common ancestor. He also *concludes* about the identity a cousin's 5th great-grandparents, breaking down a "brick wall." It is not a perfect example of triangulation. Yet, a reasonably exhaustive search has been executed, available evidence has been integrated, and no conflicts remain. One might wish for (and eventually find) additional records (atDNA matches, or otherwise) that will bolster (or debunk) the *conclusion*, but there is a *conclusion* that is viable now—a *conclusion* the author accepts now. The information bears up under scrutiny and leaves the author confident that the *conclusion* will stand as additional *information* becomes available.

PROOF EXPLAINED

Proof exists only if the *conclusion* has been recorded for others to examine. This documentation should include the matches (m) involved, the lineages relating the individuals to the common ancestor (a), and details of the testing (t) applied. In presenting results, areas of doubt should be highlighted and reasons for acceptance explained. Conflicts encountered and their resolutions must be explained. Appendices B, C, D and F document *conclusions* (or not)—proof explained.

Standards for presenting this documentation are not well developed. An area of particular interest is how to reference genotype information. The Genetic Genealogy Standards Committee promises to provide guidance on this issue.[267]

SUMMARY

This text presents a methodology for identifying, evaluating, and presenting a *conclusion* involving atDNA *evidence*. It places this methodology firmly within the framework of the genealogical research process—question asking, information gathering, hypothesis testing, conclusion accepting, and proof explained—rooting genetic genealogical practice within the processes that lead to genealogical proof.

This text focuses particularly on *hypothesis* testing as the means of evaluating the strengths and weaknesses of a *conclusion* based on atDNA *evidence*, including important heuristics that define the capabilities and limits that accompany the use of the atDNA record. Testing scrutinizes and refines *hypotheses*, ultimately making it possible to confidently use atDNA to confirm genealogical relationships.

Not all possible topics have been discussed. Instead, this text has focused on core topics used regularly in genetic genealogy (e.g., triangulation).

Genetic genealogical practice continues to evolve. New technologies and more detailed information (e.g., next-generation sequence data) are already working their way toward broad availability. Scientists continue to explain behaviors and refine heuristics that both enable and limit the use of atDNA for genealogical purposes. Even so, the methodology presented herein, and the framework within which it resides, will continue to be relevant as the genetic genealogical community of practice transforms to take advantage of these developments.

BIBLIOGRAPHY

PRIMARY SOURCES

23andMe, Inc. DNA Relatives [of GT999]: aggregate data. https://you.23andme.com/tools/relatives/ : accessed 10 June 2016.

Adjutant General's Office. Register of Enlistments in the U.S. Army, 1798-1914. SHAW, Hazael. 21 August 1812. Record Group 94: Records of the Adjutant General's Office, 1762 - 1984. Microfilm Publication M233. National Archives, Washington, DC. Collection: U.S. Army, Register of Enlistments, 1798-1914. http://search.ancestry.com/search/db.aspx?dbid=1198 : accessed 15 April 2016.

Ancestry.com. Town Clerks´ Registers of Men Who Served in the Civil War, ca 1861-1865. SHAW, Elkanah. Collection: New York, Town Clerks' Registers of Men Who Served in the Civil War, ca 1861-1865. http://search.ancestry.com/search/db.aspx?dbid=1964 : accessed 8 April 2016.

AncestryDNA™. *AncestryDNA™ Results for an8181*. http://dna.ancestry.com/ : accessed 22 January 2016.

AncestryDNA™. *AncestryDNA™ Results for [GT999]*. http://dna.ancestry.com/ : accessed 17 May 2016.

Births (CR) United States. Rochester, Windsor, Vermont. 6 September 1799. SHAW, Ruel. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1971-27949-15341-87?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > Image 74 of 368).

Burr, David H. (1829) Map of the County of St. Lawrence. New York: D H Burr. http://www.davidrumsey.com/maps4187.html : accessed 15 April 2016.

Census. 1810. United States. Rochester, Windsor, Vermont. p. 545. Collection: 1810 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7613 : accessed 3 April 2016.

Census. 1810. United States. Rochester, Windsor, Vermont. p. 546. Collection: 1810 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7613 : accessed 3 April 2016.

Census. 1820. United States. Potsdam, St Lawrence, New York. p. 62. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 7 April 2016.

Census. 1820. United States. Potsdam, St Lawrence, New York. p. 63. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 3 April 2016.

Census. 1830. United States. Potsdam, St Lawrence, New York. p. 134. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058 : accessed 7 April 2016.

Census. 1830. United States. Potsdam, St Lawrence, New York. p. 135. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058 : accessed 7 April 2016.

Census. 1830. United States. Potsdam, St Lawrence, New York. p. 151. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058

: accessed 7 April 2016.

Census. 1840. United States. Potsdam, St Lawrence, New York. p. 199. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

Census. 1840. United States. Potsdam, St Lawrence, New York. p. 209. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

Census. 1840. United States. Potsdam, St Lawrence, New York. p. 214. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

Census. 1850. United States. Dickinson, Franklin, New York. Schedule: Mortality. p. 1. Collection: New York, U.S. Census Mortality Schedules, 1850-1880. http://search.ancestry.com/search/db.aspx?dbid=1626 : accessed 17 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. p. 5B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. p. 10A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. p. 54A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 7 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. p. 54B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. p. 56A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. p. 59B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

Census. 1850. United States. Potsdam, St Lawrence, New York. Schedule: Agriculture. p. 531. Collection: Selected U.S. Federal Census Non-Population Schedules, 1850-1880. http://search.ancestry.com/search/db.aspx?dbid=1276 : accessed 7 April 2016.

Census. 1860. United States. Potsdam, St Lawrence, New York. p. 68. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

Census. 1860. United States. Potsdam, St Lawrence, New York. p. 781. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

Census. 1860. United States. Potsdam, St Lawrence, New York. p. 820. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

Census. 1900. United States. Potsdam, St Lawrence, New York. p. 10A. Collection: 1900 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7602

: accessed 21 April 2016.

Census. 1900. United States. Tracy, Lyon, Minnesota. p. 10B. Collection: 1900 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7602 : accessed 21 April 2016.

Death announcements. (1895) *Ogdensburg advance and St. Lawrence weekly Democrat.* MAYON[sic], Cynthia. 28 February. p. 1b. http://nyshistoricnewspapers.org/lccn/sn83031423/1895-02-28/ed-1/seq-1/ : accessed 20 April 2016.

Deaths (CR) United States. Carver, Plymouth, Massachusetts. 18 June 1781. SHAW, Waitstill. Collection: Massachusetts, Town and Vital Records, 1620-1988. http://search.ancestry.com/search/db.aspx?dbid=2495 : accessed 7 April 2016.

Deaths (CR) United States. Rochester, Windsor, Vermont. 4 May 1804. SHAW, Mary. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1971-27949-15341-87?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > Image 74 of 368).

Deaths (CR) United States. Rochester, Windsor, Vermont. 10 August 1803. SHAW, Freeman. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1971-27949-15341-87?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > Image 74 of 368).

Deaths (CR) United States. St Lawrence, New York. 23 December 1849. EDGERTON, Charles. Source film no: 556598. Item 2. Collection: Births, marriages and deaths, St. Lawrence County, New York, 1847-1849. https://familysearch.org/ark:/61903/3:1:3Q9M-C9BF-Q94L-5 : accessed 7 April 2016.

Deaths (CR) United States. Winona, Winona, Minnesota. 24 February 1913. CLARK, Eliza Rebecca. Source film no.: 2138535. Cert no.: 15696. Collection: Minnesota Deaths and Burials, 1835-1990. https://familysearch.org/ark:/61903/1:1:FD9Y-J9W : accessed 3 April 2016.

Department of Veterans Affairs. Revolutionary War Pension and Bounty-Land Warrant Application Files. SHAW, Daniel. Pension No.: W. 2447. National Archives, Washington, DC. Collection: Revolutionary War Pensions. https://www.fold3.com/image/20162877 : accessed 3 April 2016.

GEDmatch.com. Result of default ‘GEDmatch DNA Segment Search’ for GT999, PGT999P1 and PGT999M1. https://www.gedmatch.com/ : accessed 17 May 2016.

GEDmatch.com. Result of default ‘Matching Segment Search’ for GT712. https://www.gedmatch.com/ : accessed 23 May 2016.

GEDmatch.com. Result of default ‘Matching Segment Search’ for GT999. https://www.gedmatch.com/ : accessed 23 May 2016.

GEDmatch.com. Result of default multi-kit ‘2-D Chromosome Browser comparison’ and ‘Autosomal Matrix comparison’ between PGT368P1 and each of GT832, GT365, GT431, GT651, GT463, GT418, GT658, GT588, GT193, GT237, GT833, GT779, GT176, GT385, GT589, GT631, GT726, GT971, GT192, GT310, GT132, GT112, GT391, GT918, GT594, GT259, GT575, GT111, GT505, GT534, GT432 and GT572. https://www.gedmatch.com/ : accessed 14 May 2016.

GEDmatch.com. Result of default ‘one-to-many’ comparison for GT381. https://www.gedmatch.com/ : accessed 9 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT186 and each of GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT186 and each of GT381, GT625 and GT177. https://www.gedmatch.com/ : accessed 9 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT186 and each of GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 9 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT186 and GT136. https://www.gedmatch.com/ : accessed 9 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT383 and each of GT480 and GT654. https://www.gedmatch.com/ : accessed 1 April 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT436 and GT122. https://www.gedmatch.com/ : accessed 10 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT439 and GT557. https://www.gedmatch.com/ : accessed 7 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT480 and GT654. https://www.gedmatch.com/ : accessed 1 April 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT610 and GT926. https://www.gedmatch.com/ : accessed 14 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and each of GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and each of GT383, GT480 and GT654. https://www.gedmatch.com/ : accessed 1 April 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and each of GT610, GT116, GT341, GT654, GT383, GT480, GT938, GT196, GT557, GT554, GT338, GT369, GT728, GT167, GT820, GT633, GT966. https://www.gedmatch.com/ : accessed 17 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and each of GT709, GT610 and GT926. https://www.gedmatch.com/ : accessed 14 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT439. https://www.gedmatch.com/ : accessed 7 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT557. https://www.gedmatch.com/ : accessed 7 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT880. https://www.gedmatch.com/ : accessed 28 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT654 and each of GT611, GT610, GT116, GT341, GT383, GT480, GT938, GT196, GT557, GT554, GT338, GT369, GT728, GT167, GT820, GT633, GT966. https://www.gedmatch.com/ : accessed 17 June 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT709 and each of GT610 and GT926. https://www.gedmatch.com/ : accessed 14 May 2016.

GEDmatch.com. Result of default ‘one-to-one’ comparison between GT831 and GT124.

https://www.gedmatch.com/ : accessed 7 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparison between GT999 and GT124. https://www.gedmatch.com/ : accessed 7 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparison between GT999 and GT793. https://www.gedmatch.com/ : accessed 17 June 2016.

GEDmatch.com. Result of default 'one-to-one' comparison between GT999 and GT831. https://www.gedmatch.com/ : accessed 7 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparison between PGT999M1 and GT611, GT709, GT610, GT861, GT901, GT553, GT439 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

GEDmatch.com. Result of default 'one-to-one' comparison between PGT999M1 and GT611, GT861, GT709, GT610, GT901, GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT118, GT163, GT999 and GT186 with GT208. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT118, GT163, GT999 and GT186 with GT293. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT118, GT163, GT999 and GT186 with GT596. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT241, GT787, GT502 and GT357 with each of GT163, GT978 and GT625. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT368, GT859 and GT929 with GT136. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT368, GT859 and GT929 with GT177. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT368, GT859 and GT929 with GT978. https://www.gedmatch.com/ : accessed 20 May 2016.

GEDmatch.com. Result of default X 'one-to-one' comparison between GT999 and GT793. https://www.gedmatch.com/ : accessed 17 June 2016.

GEDmatch.com. Result of 'Multiple Kit Analysis: Autosomal DNA comparison matrix' for GT611, GT610, GT116, GT341, GT654, GT383, GT480, GT938, GT196, GT557, GT554, GT338, GT369, GT728, GT167, GT820, GT633, GT966. https://www.gedmatch.com/ : accessed 13 May 2016.

GEDmatch.com. Result of 'Multiple Kit Analysis: Autosomal DNA comparison matrix' for GT999, GT163, GT625, GT177, GT381, GT136, GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 13 May 2016.

GEDmatch.com. Result of 'Multiple Kit Analysis: Autosomal DNA comparison matrix' for GT999, GT611, GT709, GT610, GT861, GT901, GT553, GT439 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

GEDmatch.com. Result of ‘Multiple Kit Analysis: Autosomal DNA comparison matrix’ for GT999, GT611, GT709, GT610, GT861, GT901, GT553, GT439, GT122 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

GEDmatch.com. Result of ‘Multiple Kit Analysis: Autosomal DNA comparison matrix’ for GT999, GT611, GT861, GT709, GT610, GT901, GT124, GT732, GT681, and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

GEDmatch.com. Result of ‘one-to-one’ comparison (using 300 SNP and 4 cM match thresholds) between GT611 and each of GT654, GT480 and GT383. https://www.gedmatch.com/ : accessed 16 June 2016.

GEDmatch.com. Result of ‘one-to-one’ comparison (using 300 SNP and 4 cM match thresholds) between PGT999M1 and each of GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

GEDmatch.com. Result of ‘one-to-one’ comparison (using 300 SNP and 4 cM match thresholds) between PGT999P1 and each of GT163, GT625, GT177, GT381, GT136, GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 1 April 2016.

Green Mountain Patriot (Peacham, Vermont). (1801) Justices of the Peace. Green Mountain Patriot. 12 November. p. 3c. http://www.genealogybank.com/ : accessed 21 February 2016.

Jackson, Ronald V & Accelerated Indexing Systems. NY 1815 Port Arrivals. SHAW. 1815. Collection: New York, Compiled Census and Census Substitutes Index, 1790-1890. http://search.ancestry.com/search/db.aspx?dbid=3564 : accessed 3 April 2016.

Marriages (CR) United States. Northfield, Washington, Vermont. 12 November 1873. MORGAN, Joseph L and HENRY, Katherine A. Collection: Vermont, Vital Records, 1720-1908. http://search.ancestry.com/search/db.aspx?dbid=4661 : accessed 3 April 2016.

Marriages (CR) United States. Plympton, Plymouth, Massachusetts. 6 August 1778. SHAW, Daniel and BARROWS, Mary. Collection: Massachusetts, Town and Vital Records, 1620-1988. http://search.ancestry.com/search/db.aspx?dbid=2495 : accessed 3 April 2016.

Marriages (CR) United States. Rochester, Windsor, Vermont. 18 March 1804. SHAW, Daniel Junr and AUSTIN, Sally. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1942-27949-15669-97?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > image 73 of 368).

Monumental inscriptions. United States. Bayside Cemetery, Potsdam, St Lawrence, New York. 10 December 1860. CHANDLER, Waitstill Shaw. Transcribed by gravehunter1218. Find A Grave Memorial: 75463137. http://www.findagrave.com/ : accessed 7 April 2016.

Monumental inscriptions. United States. Bayside Cemetery, Potsdam, St Lawrence, New York. 18 March 1855. TRAVER, Susan. Transcribed by gravehunter1218. Find A Grave Memorial: 148278717. http://www.findagrave.com/ : accessed 3 April 2016.

Monumental inscriptions. United States. Lakenham Cemetery, Carver, Plymouth, Massachusetts. 18 June 1781. SHAW, Waitstill. Transcribed by Anne Shurtleff Stevens. Find A Grave Memorial: 36893709. http://www.findagrave.com/ : accessed 7 April 2016.

Monumental inscriptions. United States. North Bridgewater Cemetery, Bridgewater, Windsor, Vermont. 15 August 1850. SHAW, Elkanah. Transcribed by David Edsall. Find A Grave Memorial: 111849132. http://www.findagrave.com/ : accessed 17 April 2016.

Monumental inscriptions. United States. Union Cemetery, Norwood, St Lawrence, New York. 20 December 1865. EDGERTON, Elizabeth Shaw. Transcribed by Anne Cady. Find A Grave Memorial: 41185980. http://www.findagrave.com/ : accessed 3 April 2016.

Monumental inscriptions. United States. Union Cemetery, Norwood, St Lawrence, New York. 22 December 1834. SHAW, Sally. Transcribed by Anne Cady. Find A Grave Memorial: 41186042. http://www.findagrave.com/ : accessed 3 April 2016.

Monumental inscriptions. United States. Union Cemetery, Norwood, St Lawrence, New York. 22 March 1844. SHAW, Daniel. Transcribed by Anne Cady. Find A Grave Memorial: 41186034. http://www.findagrave.com/ : accessed 3 April 2016.

Notice of probate. (1846) *St. Lawrence Republican*. CHANDLER, John. 28 July. p. 4d. http://nyshistoricnewspapers.org/lccn/sn83031401/1846-07-28/ed-1/seq-4 : accessed 15 April 2016.

Obituaries. (1935) Herald-Recorder (Potsdam, NY). 19 April. BURTON, Herbert S. p. 4d. http://nyshistoricnewspapers.org/lccn/sn84035824/1935-04-19/ed-1/seq-4/ : accessed 10 April 2016.

Potsdam Herald-Recorder. (1934) Speaks on 'The Union'. *The Potsdam Herald-Recorder.* 14 September. p. 1e, 3c&d. http://nyshistoricnewspapers.org/lccn/sn84035824/1934-09-14/ed-1/seq-1 : accessed 17 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. CHANDLER, John. 30 October 1811. p. 31, 47, 61, 85. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. EDGERTON, Charles. 19 September 1812. p. 43. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. EDGERTON, Charles. 30 November 1816. p. 71. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. MORGAN, Forest. 30 November 1816. p. 51, 71. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. MORGAN, Forest. 30 November 1818. p. 91. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Daniel. 30 November 1816. p. 51, 71. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Daniel Jr. 30 November 1818. p. 73. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Daniel; SHAW, Elkanah; SHAW, Daniel Jr; SHAW, Salmon; SHAW, Hazael; MORGAN, Forest; CHANDLER, John; and EDGERTON, Charles. 30 November 1810. p. 27, 31, 43, 45, 47, 49, 51, 59, 61, 63, 71, 83, 85, 87, 91. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Elkanah. 30 November 1810. p. 27. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Elkanah. 30 November 1817. p. 63, 87. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Hazael. 23 August 1817. p. 63, 87. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Salmon. 20 September 1817. p. 49. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Salmon. 30 November 1810. p. 27, 45, 59, 83. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

Rogerson, Andrew E. (1858) Map of St. Lawrence Co. New York. Philadelphia: J.B. Shields Publisher. https://www.loc.gov/item/2006626022/ : accessed 15 April 2016.

Spooner's Vermont Journal (Windsor, Vermont). (1812) List of Letters. Spooner's Vermont Journal. 6 January. p. 3d; 13 January. p. 3b; 20 January. p. 4c; 27 January. p. 4c; 10 February. p. 4b. http://www.genealogybank.com/ : accessed 21 February 2016.

St Lawrence Herald (Potsdam, NY). (1903) Pioneers of Postdam: A List of People Who

First Settled the Town. The St Lawrence Herald. 22 May. p. 1e&f. http://www.genealogybank.com/ : accessed 8 April 2016.

St Lawrence Plaindealer (Canton, NY). (1937) Clarks Crossing: An Historic Spot. *St Lawrence Plaindealer*. 28 September. p. 7b&c. http://fultonhistory.com/ : accessed 2 April 2016.

Testamentary records. United States. 4 January 1850. EDGERTON, Charles. Administration. Collection: New York, Wills and Probate Records, 1659-1999. http://search.ancestry.com/search/db.aspx?dbid=8800 : accessed 17 April 2016.

Testamentary records. United States. 13 December 1855. EDGERTON, Ransom G. Administration. Collection: New York, Wills and Probate Records, 1659-1999. http://search.ancestry.com/search/db.aspx?dbid=8800 : accessed 17 April 2016.

Testamentary records. United States. 23 February 1893. SHAW, Daniel. Notice of Probate. Collection: New York, Wills and Probate Records, 1659-1999. http://search.ancestry.com/search/db.aspx?dbid=8800 : accessed 20 April 2016.

Vermont Republican (Windsor, Vermont). (1810) Legislature of Vermont: Appointments of Justice of the Peace. Vermont Republican. 29 October. p. 6a. http://www.genealogybank.com/ : accessed 21 February 2016.

SECONDARY SOURCES

23andMe, Inc. *23andMe - DNA Genetic Testing & Analysis*. https://www.23andme.com : accessed 3 June 2016.

23andMe, Inc. *23 And Me TV Commercial*. https://youtu.be/LrtPoke4X2g : accessed 30 January 2016.

an8181. *[an8181's] Biological Family Tree*. http://trees.ancestry.com/tree/86926265/family : accessed 25 December 2015.

[an8181's cousin]. *[an8181's cousin] Family Tree*. http://trees.ancestry.com/tree/9335040/family : accessed 23 January 2016.

Ancestry.com. *Ancestry.com Launches New AncestryDNA Service: The Next Generation of DNA Science Poised to Enrich Family History Research: Affordable DNA Test Combines Depth of Ancestry.com Family History Database With an Extensive Collection of DNA Samples to Open New Doors to Family Discovery*. http://ir.ancestry.com/releasedetail.cfm?ReleaseID=669964 : accessed 12 April 2016.

Ancestry.com. *Public Member Trees*. http://search.ancestry.com/search/db.aspx?dbid=1030 : accessed 6 December 2015.

AncestryDNA™. (2014) *Your AncestryDNA results are in!*. E-mail to Thad Thomas. 4:13 AM. 6 April. info@blueskiff.com.

AncestryDNA™. (2015) *Your AncestryDNA results are in!*. E-mail to Thad Thomas. 11:12 AM. 8 April. info@blueskiff.com.

AncestryDNA™. *Ancestry.com DNA Kit TV Commercial, 'Family History'*. http://ispot.tv/a/7g1I : accessed 30 January 2016.

AncestryDNA™. *Ancestry.com DNA TV Commercial, '1000 Years in the Past'*. http://ispot.tv/a/7CYh : accessed 30 January 2016.

AncestryDNA™. *AncestryDNA Matching Help and Tips: Should other family members get tested?* http://dna.ancestry.com/ : accessed 4 February 2016.

AncestryDNA™. *AncestryDNA TV Commercial, 'Lederhosen'*. http://ispot.tv/a/7c4Y : accessed 30 January 2016.

AncestryHealth. *Family Health History*. https://health.ancestry.com : accessed 3 June 2016.

Ballard, Elizabeth Wilson. *Kenny, Kenny, Kenny…*. https://digginupgraves.wordpress.com/2015/04/10/kenny-kenny-kenny/ : accessed 13 March 2016.

Ball, Catherine A, Barber, Mathew J, Byrnes, Jake, et al. (2016) *AncestryDNA^{TM} Matching White Paper: Discovering genetic matches across a massive, expanding genetic database*. Lehi, Utah: AncestryDNA^{TM}. http://dna.ancestry.com/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf : accessed 17 May 2016.

Ball, Catherine A, Barber, Mathew J, Byrnes, Jake K, et al. (2014) *DNA Circles^{TM} White Paper: Identifying groups of descendants using pedigrees and genetically inferred relationships in a large database*. Lehi, Utah: AncestryDNA^{TM}. http://dna.ancestry.com/resource/whitePaper/AncestryDNA-DNA-Circles-White-Paper : accessed 26 May 2016.

Bartlett, Jim. *segment-ology*. http://segmentology.org/ : accessed 17 November 2015.

Belluz, Julia. *Genetic testing brings families together*. http://www.vox.com/2014/9/9/6107039/23andme-ancestry-dna-testing : accessed 18 March 2016.

Belluz, Julia. *With genetic testing, ['George Doe'] gave [his] parents the gift of divorce*. http://www.vox.com/2014/9/9/5975653/with-genetic-testing-i-gave-my-parents-the-gift-of-divorce-23andme : accessed 18 March 2016.

Bettinger, Blaine. *The Genetic Genealogist: Adding DNA to the Genealogist's Toolbox*. http://thegeneticgenealogist.com/ : accessed 24 November 2015.

Bettinger, Blaine T. *The Recombination Project: Analyzing Recombination Frequencies Using Crowdsourced Data*. https://thegeneticgenealogist.com/wp-content/uploads/2017/02/Recombination_Preprint.pdf : accessed 2 December 2017.

Blouin, Michael S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*. 18 (10). pp. 503–511. http://www.sciencedirect.com/science/article/pii/S0169534703002258 : accessed 14 February 2016.

Board for Certification of Genealogists. *About BCG*. http://www.bcgcertification.org/aboutbcg/index.html : accessed 28 January 2016.

Board for Certification of Genealogists. (2014) *Genealogy Standards*. 50th Anniversary Ed. Nashville, Tennessee: Ancestry.com.

Board for Certification of Genealogists. (2000) *The BCG Genealogical Standards Manual*. Nashville, Tennessee: Ancestry Publishing.

Browning, Brian L. & Browning, Sharon R. (2011) A Fast, Powerful Method for Detecting Identity by Descent. *American Journal of Human Genetics*. 88 (2). pp. 173–182. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035716/ : accessed 14 February 2016.

Browning, Brian L. & Browning, Sharon R. (2013) Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics*. 194 (2). pp. 459–471. http://www.genetics.org/content/194/2/459 : accessed 14 February 2016.

Browning, Sharon R. & Browning, Brian L. (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*. 12 (10). pp. 703–714. http://www.nature.com/doifinder/10.1038/nrg3054 : accessed 20 February 2016.

Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

Bryant, Rebecca. *A Practical Guide to Using Autosomal DNA (atDNA) for Genealogical Purposes*. https://sites.google.com/site/bryantsofrockislandcreek/2-european-heritage/dna-results/autosomal-dna-atdna : accessed 4 November 2015.

Campbell, Christopher L., Furlotte, Nicholas A., Eriksson, Nick, et al. (2015) Escape from crossover interference increases with maternal age. *Nature Communications*. 6 p. 6260.

Canada, R A. *How does Population Finder determine the percentages of different ancestries?* https://web.archive.org/web/20140606004944/https://www.familytreedna.com/learn/autosomal-ancestry/ethnic-origins/population-finder-find-ancestry/ : accessed 4 June 2016.

Captain, Sean. *Biotech Firms Battle Over Who Owns Genetic Data*. http://www.fastcompany.com/3057525/biotech-firms-battle-over-who-owns-genetic-data : accessed 31 March 2016.

Carmi, Shai, Palamara, Pier Francesco, Vacic, Vladimir, et al. (2013) The Variance of Identity-by-Descent Sharing in the Wright–Fisher Model. *Genetics*. 193 (3). pp. 911–928. http://www.genetics.org/content/193/3/911 : accessed 16 February 2016.

Chen, Daphne. *Should scientists be allowed to change DNA to prevent genetic disease?* https://www.ksl.com/?sid=38376617 : accessed 4 February 2106.

Coble, Michael D., Loreille, Odile M., Wadhams, Mark J., et al. (2009) Mystery Solved: The Identification of the Two Missing Romanov Children Using DNA Analysis. *PLOS ONE*. 4 (3). p. e4838. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004838 : accessed 5 March 2016.

Columbia University and New York Genome Center. *DNA Land*. https://dna.land/ : accessed 3 February 2016.

Coop, Graham, Wen, Xiaoquan, Ober, Carole, et al. (2008) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science*. 319 (5868). pp. 1395–1398.

Cooper, Kitty Munson. *Kitty Cooper's Blog*. http://blog.kittycooper.com/ : accessed 10 December 2015.

Cooper, Kitty Munson. (2015) 'RT3214 - How to do a DNA Triangulation: Case Studies.' *RootsTech 2016*. Salt Lake City 6 February 2016. http://www.rootstech.org/about/syllabus?lang=eng : accessed 1 February 2016.

Coop, Graham. *gcbias*. https://gcbias.org/ : accessed 4 February 2016.

Cowan, Crista. (2016) 'DNA Test Results: Handling the Unexpected.' *RootsTech 2016*. Salt Lake City 6 February 2016.

Custer, Nancy V. *Meiosis: How chromosomes are passed from parent to offspring*. http://www.contexo.info/DNA_Basics/Meiosis.htm : accessed 25 October 2015.

Dallas, Kelsey. (2016) Finding God in one of science's biggest debates — genetic editing. *Deseret News*. 13 January. http://national.deseretnews.com/article/16536/finding-god-in-one-of-sciences-biggest-debates-genetic-editing.html : accessed 14 January 2016.

Dictionary.com. (n.d.) *Dictionary.com Unabridged.* Random House, Inc. http://www.dictionary.com/ : accessed 2 June 2016.

DNAAdoption, Inc. (2014) *Using GEDmatch*. http://moodle.dnagedcom.com/MoodleClass/DNAadoption/UsingGEDMATCHRevised3-2014.pdf : accessed 31 October 2015.

DNA.land. *Face it: DNA cannot find all your relatives*. https://medium.com/@dl1dl1/face-it-dna-cannot-find-all-your-relatives-f68089b8e1e9#.squu4hdfv : accessed 3 March 2016.

Donnelly, Kevin P. (1983) The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*. 23 (1). pp. 34–63.

Dudley, Joel T. & Karczewski, Konrad J. (2013) *Exploring Personal Genomics*. Oxford University Press. http://www.oxfordscholarship.com/ : accessed 16 February 2016.

Durand, Eric Y., Eriksson, Nicholas, & McLean, Cory Y. (2014) Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution*. 31 (8). pp. 2212–2222. http://mbe.oxfordjournals.org/content/31/8/2212 : accessed 27 November 2015.

Durant, Samuel W. (1878) *History of St. Lawrence Co., New York : illustrations and biographical sketches, some of its prominent men and pioneers*. Philadelphia, Pennsylvania: L.H. Everts. http://www.ancestry.com/ : accessed 16 April 2016.

Dziebel, German. *The Effect of Long-Term Endogamy on Identity-By-Descent*. http://anthropogenesis.kinshipstudies.org/2012/04/the-effect-of-long-term-endogamy-and-cryptic-consanguinity-on-identity-by-descent/ : accessed 26 February 2016.

Eccles Institute of Human Genetics (University of Utah). *Human Genetics*. http://genetics.utah.edu/ : accessed 10 December 2015.

Edwards, Dan. *Chromosome Pile-Ups in Genetic Genealogy: Examples from 23andMe and FTDNA*. http://ourpuzzlingpast.com/geneblog/2015/01/31/chromosome-pile-ups-in-genetic-genealogy-examples-from-23andme-and-ftdna/ : accessed 23 May 2016.

Estes, Roberta. *DNAeXplained – Genetic Genealogy*. http://dna-explained.com/ : accessed 9 January 2016.

Estes, Roberta & Corbeil, Karin. *Demystifying Ancestry's Relationship Predictions Inspires New Relationship Estimator Tool*. http://dna-explained.com/2016/02/22/demystifying-ancestrys-relationship-predictions-inspires-new-relationship-estimator-tool/ : accessed 22 February 2016.

Evans, James P. (2008) Recreational genomics; what's in it for you? *Genetics in Medicine*. 10 (10). pp. 709–710. http://www.nature.com/gim/journal/v10/n10/full/gim2008108a.html : accessed 30 January 2016.

FamilySearch International. *Family Tree*. https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 6 December 2015.

FamilySearch International. *GEDCOM X and the Genealogical Research Process*. http://www.gedcomx.org/GEDCOM-X-and-the-Genealogical-Research-Process.html : accessed 21 April 2016.

FamilySearch Wiki. *Upcoming Conferences*. https://familysearch.org/learn/wiki/en/Upcoming_Conferences : accessed 30 January 2016.

Foard, Mesa. (2014) *A Methodology: Identifying your Relatives through your atDNA Results*. http://moodle.dnagedcom.com/MoodleClass/General/Methodology_revised.pdf : accessed 31 October 2015.

Foster, Eugene A., Jobling, M. A., Taylor, P. G., et al. (1998) Jefferson fathered slave's last child. *Nature*. 396 (6706). pp. 27–28. http://www.nature.com/nature/journal/v396/n6706/full/396027a0.html : accessed 5 March 2016.

Frisbie, Richard. *Vermont/New York Boundary History*. http://www.hopefarm.com/vermont.htm : accessed 21 April 2016.

Full Genomes Corporation, Inc. *Home*. https://www.fullgenomes.com : accessed 31 May 2016.

Gannon, Megan. *It's Really Richard: DNA Confirms King's Remains*. http://www.livescience.com/48963-king-richard-iii-dna-confirmed.html : accessed 5 March 2016.

GEDCOM X Project. *GEDCOM X and the Genealogical Research Process*. http://www.gedcomx.org/GEDCOM-X-and-the-Genealogical-Research-Process.html : accessed 5 March 2016.

GEDmatch.com. *GEDmatch: Tools for DNA and Genealogy Research*. https://www.gedmatch.com/ : accessed 4 June 2016.

Gene By Gene, Ltd. *Autosomal Tests*. https://www.familytreedna.com/learn/dna-basics/autosomal/ : accessed 4 June 2016.

Genes and Trees. (no date) 'GEDMatch Basics.' https://www.youtube.com/watch?v=acGJmLlsWg4&feature=youtu.be : accessed 1 November 2015.

Genetic Genealogy Standards Committee. *Genetic Genealogy Standards*. http://www.geneticgenealogystandards.com : accessed 15 February 2106.

Genome Research Limited. *Personal genomics: the future of healthcare?* http://www.yourgenome.org/stories/personal-genomics-the-future-of-healthcare : accessed 30 January 2016.

Gleeson, Maurice. *DNA and Family Tree Research: A Systematic Approach to Analysing your Autosomal DNA Matches - Introduction*. http://dnaandfamilytreeresearch.blogspot.com/2013/05/a-systematic-approach-to-analysing-your.html : accessed 13 May 2016.

Greenawalt, Lindsay M. *The DNA numbers game*. http://cryokidconfessions.blogspot.com/2011/09/dna-numbers-game.html : accessed 6

December 2015.

Griffith, Sue. *DNA Spreadsheets!*. http://www.genealogyjunkie.net/dna-spreadsheets.html : accessed 2 November 2015.

Griffith, Sue. *DNA Tips, Tools, & Managing Matches*. http://www.genealogyjunkie.net/dna-tips-tools--managing-matches.html : accessed 2 November 2015.

Griffith, Sue. *Miscellaneous Notes On DNA Matches And Genetic Genealogy*. http://www.genealogyjunkie.net/miscellaneous-notes.html : accessed 2 November 2015.

[GT124]. *[GT214] Family Tree*. http://trees.ancestry.com/tree/80075627/family : accessed 4 January 2016.

[GT177]. *Family Research tree (DNA matches)*. http://person.ancestry.com/tree/81972480/person/74006448666 : accessed 10 June 2016.

[GT357]. (2016a) *Re: DNA Match*. E-mail to Thad Thomas. 4:08 AM. 3 May. thad.thomas.2013@uni.strath.ac.uk.

[GT357]. (2016b) *Re: DNA Match*. E-mail to Thad Thomas. 8:32 PM. 14 May. thad.thomas.2013@uni.strath.ac.uk.

[GT357]. (2016c) *Re: DNA Match*. E-mail to Thad Thomas. 9:56 PM. 25 April. thad.thomas.2013@uni.strath.ac.uk.

[GT381]. (2016) *Re: MRCA: Joseph Call & Mary Sanderson*. E-mail to Thad Thomas. 6:13 AM. 30 April. thad.thomas.2013@uni.strath.ac.uk.

[GT383]. *Ancestors of [GT383]*. http://freepages.folklore.rootsweb.ancestry.com/~fromheretopaternity/index.htm : accessed 1 April 2016.

[GT385 administrator]. (2016) *Re: Looking for common ancestor with [GT385] ...*. E-mail to Thad Thomas. 06:21 PM. 8 March. thad.thomas.2013@uni.strath.ac.uk.

[GT436]. *[GT436] Family Tree*. http://person.ancestry.com/tree/243923/person/-2104873513 : accessed 27 December 2015.

[GT439]. *[GT439] Family Tree*. http://trees.ancestry.com/tree/16474504/family : accessed 25 December 2015.

[GT439]. (2015a) *Re: My CA with Gladys?* E-mail to Thad Thomas. 11:09 PM. 27 December. thad.thomas.2013@uni.strath.ac.uk.

[GT439]. (2015b) *Spreadsheet Template*. E-mail to Thad Thomas. 9:17 PM. 11 November. thad.thomas.2013@uni.strath.ac.uk.

[GT480]. *[GT480 Family Tree]*. http://trees.ancestry.com/tree/70185236/family : accessed 1 April 2016.

[GT594]. *[GT594] Family Tree*. http://trees.ancestry.com/tree/56467638/family : accessed 7 April 2016.

[GT654]. *[GT654] Family Tree*. http://trees.ancestry.com/tree/63112612/family : accessed 1 April 2016.

[GT779]. (2016a) *RE: Looking for common ancestor with [GT779] ...*. E-mail to Thad Thomas. 08:05 AM. 8 March. thad.thomas.2013@uni.strath.ac.uk.

[GT779]. (2016b) *RE: Looking for common ancestor with [GT779] ...*. E-mail to Thad Thomas. 12:33 PM. 7 March. thad.thomas.2013@uni.strath.ac.uk.

[GT793]. (2016) *Re: DNA cousin on GEDmatch*. E-mail to Thad Thomas. 02:37 AM. 21 January. thad.thomas.2013@uni.strath.ac.uk.

[GT831]. *[GT831] Family Tree*. http://trees.ancestry.com/tree/15002490/family : accessed 25 December 2015.

[GT880]. *[GT880]lineagetree*. http://person.ancestry.com/tree/1825910/person/-1753013398 : accessed 28 May 2016.

[GT978]. *[GT978] Family Tree*. http://trees.ancestry.com/tree/51133039/person/13271713048 : accessed 25 April 2016.

[GT978]. (2016) *Re: Cyril Call and Sally Tiffany*. E-mail to Thad Thomas. 5:11 PM. 25 April. thad.thomas.2013@uni.strath.ac.uk.

[GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 25 December 2015.

[GT999]. *Zee test Family Tree*. http://trees.ancestry.com/tree/73592017/family : accessed 25 December 2015.

Guy, Louise. (2015) Re: Ghost twins. PG Genealogical, Palaeographic & Heraldic Studies General (ALL LEVELS): Persiflage Forum, 8 November. http://classes.myplace.strath.ac.uk/mod/forum/discuss.php?d=185282&postid=587769 : accessed 16 November 2015.

Harman-Hoog, Diane. *A Methodology to Identify Relatives with autosomal DNA Test Data*. http://dnaadoption.com/uploads/DNAadoption/DNAadoption_files/General/Methodology_for_Researching_Autosomal_DNA_Results_V3_1-9-2015.pdf : accessed 10 December 2015.

Hartley, Joel M. *Beware the False DNA Match – Hartley DNA & Genealogy*. http://www.jmhartley.com/HBlog/?p=540 : accessed 13 May 2016.

Henn, Brenna M., Hon, Lawrence, Macpherson, J. Michael, et al. (2012) Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PLoS ONE*. 7 (4). p. e34267. http://dx.doi.org/10.1371/journal.pone.0034267 : accessed 16 February 2016.

Hill, Kashmir. *Cops are asking Ancestry.com and 23andMe for their customers' DNA*. http://fusion.net/story/215204/law-enforcement-agencies-are-asking-ancestry-com-and-23andme-for-their-customers-dna/ : accessed 19 October 2015.

Hill, Richard. (2012) *Finding Family: My Search for Roots and the Secrets in My DNA*. Scotts Valley, California: On-Demand Publishing, LLC.

Hill, W. G. & Weir, B. S. (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*. 93 (1). pp. 47–64. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3070763/ : accessed 13 February 2016.

Hill, William G. & White, Ian M. S. (2013) Identification of Pedigree Relationship from Genome Sharing. *G3: Genes|Genomes|Genetics*. 3 (9). pp. 1553–1571.

http://www.g3journal.org/content/3/9/1553 : accessed 13 February 2016.

Hoitink, Yvette. *Three Things I Learned About DNA at WDYTYA Live*. http://www.dutchgenealogy.nl/three-things-i-learned-about-dna-at-wdytya/ : accessed 6 December 2015.

Hollands, Gareth J., French, David P., Griffin, Simon J., et al. (2016) The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. *BMJ*. 352 p. i1102. http://www.bmj.com/content/352/bmj.i1102 : accessed 30 March 2016.

Houghton Mifflin Harcourt Publishing Company. (2011) 'genetics.' *American Heritage Dictionary of the English Language*. Houghton Mifflin Harcourt Publishing Company. http://www.thefreedictionary.com/genetics : accessed 2 June 2016.

Huff, Chad D., Witherspoon, David J., Simonson, Tatum S., et al. (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research*. 21 (5). pp. 768–774. http://genome.cshlp.org/content/21/5/768 13 February 2016.

[Husband of an8181]. (2016) *Adoption Information*. E-mail to Thad Thomas. 10:55:12 AM. 23 January. thad.thomas@ldschurch.org.

Illumina, Inc. *HumanOmniExpress BeadChip Kit*. http://www.illumina.com/products/human_omni_express_beadchip_kits.html : accessed 26 March 2016.

International Society of Genetic Genealogy. *DNA-NEWBIE: About Group*. https://groups.yahoo.com/neo/groups/DNA-NEWBIE/info : accessed 30 January 2016.

International Society of Genetic Genealogy. *International Society of Genetic Genealogy (ISOGG) [Facebook Group]*. https://www.facebook.com/groups/isogg/ : accessed 30 January 2016.

ISOGG Wiki. *Wiki Welcome Page*. http://isogg.org/wiki/ : accessed 22 October 2015.

Jackson Laboratory. *The Difference Between Genetics and Genomics*. https://www.jax.org/genetics-and-healthcare/genetics-and-genomics/the-difference-between-genetics-and-genomics : accessed 2 June 2016.

Janzen, Tim. (2014a) 'Advanced Techniques for Use of Autosomal DNA Tests to Break through Genealogical Brick Walls.' *RootsTech 2014*. Salt Lake City 6 February 2014. http://tinyurl.com/lbd9xm8 : accessed 22 October 2015.

Janzen, Tim. (2014b) *Advanced Techniques for Use of Autosomal DNA Tests to Break through Genealogical Brick Walls*. http://www.broadcast.lds.org/elearning/fhd/Local_Support/RootsTech2014/Eng/outlines/RT1374Outline.pdf : accessed 27 November 2015.

Janzen, Tim. (2011) *[AUTOSOMAL-DNA] Chromosome mapping*. AUTOSOMAL-DNA, 7 June 2011 17:43:39 -0700. http://archiver.rootsweb.ancestry.com/th/read/AUTOSOMAL-DNA/2011-06/1307493819 : accessed 23 October 2015.

Janzen, Tim. (2015a) 'Autosomal DNA Chromosome Mapping Workshop.' *Southern California Genealogy Jamboree 2015*. Burbank 5 June 2015. http://tinyurl.com/pjg4akw : accessed 20 December 2015.

Janzen, Tim. (2014c) *Chromosome Mapping for Genetic Genealogy*. http://tinyurl.com/canzmsa : accessed 27 November 2015.

Janzen, Tim. (2014d) 'Discovering and Verifying your Ancestry using Family Finder.' *2014 International Conference on Genetic Genealogy*. Houston 11 October 2016. http://tinyurl.com/p22ejo4 : accessed 18 February 2016.

Janzen, Tim. (2015b) *RE: Autosomal DNA mapping*. E-mail to Thad Thomas. 8:30 AM. 28 November. thad.thomas.2013@uni.strath.ac.uk.

Janzen, Tim. (2014e) *Re: [AUTOSOMAL-DNA] Need Help on Triangulated Groups or Phasing and Chromosome Mapping*. AUTOSOMAL-DNA, 15 April 2014 21:49:02 -0700. http://archiver.rootsweb.ancestry.com/th/read/AUTOSOMAL-DNA/2014-04/1397623742 : accessed 20 May 2016.

Janzen, Tim. (2016) 'Using DNA to Solve Genealogical Problems.' *RootsTech 2016*. Salt Lake City 6 February 2016. http://tinyurl.com/zsukg8d : accessed 6 February 2016.

Janzen, Tim & Aulicino, Emily. (2013) 'Basics of Chromosome Mapping.' *GFO DNA Interest Group Meeting*,. Portland, Oregon 27 July 2013. https://dl.dropboxusercontent.com/u/21841126/Basics%20of%20Chromosome%20Mapping.docx : accessed 23 October 2015.

J Craig Venter Institute. *Genetics and Genomics Timeline*. http://www.genomenewsnetwork.org/resources/timeline/: accessed 30 May 2016.

Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC.

Jolie, Angelina. (2013) My Medical Choice by Angelina Jolie. *The New York Times*. 14 May. http://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html : accessed 11 March 2016.

Jones, Thomas W. (2013) *Mastering Genealogical Proof*. [Kindle version]. Arlington, Virginia: National Genealogical Society.

Jostins, Luke. *How Many Ancestors Share Our DNA?* http://www.genetic-inference.co.uk/blog/2009/11/how-many-ancestors-share-our-dna/ : accessed 4 February 2016.

Julie Granka. *The DNA matching research and development life cycle*. http://blogs.ancestry.com/techroots/the-dna-matching-research-and-development-life-cycle/ : accessed 4 June 2016.

Kapple, Annette Diane. *DNA: Triangulation, Pileups & Endogamy*. http://annettekapple.blogspot.com/2016/01/dna-triangulation-pileups-endogamy.html : accessed 9 January 2016.

Kennett, Debbie Cruwys. *Cruwys news*. http://cruwys.blogspot.com/ : accessed 12 March 2016.

King, Turi E., Fortes, Gloria Gonzalez, Balaresque, Patricia, et al. (2014) Identification of the remains of King Richard III. *Nature Communications*. 5 p. 5631. http://www.nature.com/ncomms/2014/141202/ncomms6631/full/ncomms6631.html : accessed 5 March 2016.

Kong, Augustine, Gudbjartsson, Daniel F., Sainz, Jesus, et al. (2002) A high-resolution recombination map of the human genome. *Nature Genetics*. 31 (3). p. 241.

Kong, Augustine, Thorleifsson, Gudmar, Gudbjartsson, Daniel F., et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*. 467 (7319). p. 1099.

Kyriazopoulou-Panagiotopoulou, Sofia, Kashef Haghighi, Dorna, Aerni, Sarah J., et al. (2011) Reconstruction of genealogical relationships with applications to Phase III of HapMap. *Bioinformatics*. 27 (13). pp. i333–i341. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117348/ : accessed 14 February 2016.

Leary, Helen F M. *Skillbuilding: Evidence Revisited: DNA, POE, and GPS*. http://www.bcgcertification.org/skillbuilders/learyevidence.html : accessed 16 November 2015.

Lee, Jason. *DNA Genealogy*. http://dnagenealogy.tumblr.com/ : accessed 22 February 2016.

Levin, A J. *Downside to DNA, Part 2*. https://thegenealogistdotca.wordpress.com/2016/03/24/downside-to-dna-part-2/ : accessed 12 April 2016.

Lewis, Ricki. *Direct-to-Consumer Genetic Testing: A New View*. http://blogs.plos.org/dnascience/2012/11/08/direct-to-consumer-genetic-testing-a-new-view/ : accessed 30 May 2016.

Li, Hong, Glusman, Gustavo, Hu, Hao, et al. (2014) Relationship Estimation from Whole-Genome Sequence Data. *PLoS Genet*. 10 (1). p. e1004144. http://dx.doi.org/10.1371/journal.pgen.1004144 : accessed 13 February 2016.

Lucy Holman Rector. (2008) Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*. 36 (1). pp. 7–22. http://www.emeraldinsight.com/doi/abs/10.1108/00907320810851998 : accessed 5 March 2016.

Lynn Grub, ed. (2015) *The Adoptee Survival Guide: Adoptees Share Their Wisdom and Tools*. Scotts Valley, California: CreateSpace.

Marwa. *Dystopia in Gattaca and Discrimination against Genes*. http://gandt.blogs.brynmawr.edu/web-papers/web-papers-3/dystopia-in-gattaca-and-discrimination-against-genes/ : accessed 11 March 2016.

Matise Laboratory of Computational Genetics. *Map Interpolator of the Rutgers Map*. http://compgen.rutgers.edu/mapinterpolator : accessed 3 June 2016.

McAllister, Marion, Moldovan, Ramona, Paneque, Milena, et al. (2016) The need to develop an evidence base for genetic counselling in Europe. *European Journal of Human Genetics*. 24 (4). pp. 504–505. http://www.nature.com/ejhg/journal/v24/n4/full/ejhg2015134a.html : accessed 30 March 2016.

Merriam-Webster.com. (n.d.) *Merriam-Webster Web*. Merriam-Webster. http://www.merriam-webster.com/dictionary/ : accessed 15 June 2016.

Merriman, Brenda Dougall. (2010) *Genealogical Standards of Evidence: A Guide for Family Historians*. Toronto: Dundurn Press.

Mills, Elizabeth Shown. (2009) *Evidence Explained: Citing History Sources from Artifacts to Cyberspace*. 2nd Ed. Baltimore, Maryland: Genealogical Publishing Company.

Moore, CeCe. *Adoption and DNA*. http://adopteddna.com : accessed 12 December 2015.

Moore, CeCe. *DNA Testing for Genealogy – Getting Started, Part Four*. http://www.geni.com/blog/dna-testing-for-genealogy-getting-started-part-four-376433.html : accessed 18 March 2016.

Moore, CeCe. *DNA Testing for Genealogy – Getting Started, Part One*. http://www.geni.com/blog/dna-testing-for-genealogy-getting-started-part-one-375984.html : accessed 18 March 2016.

Moore, CeCe. *DNA Testing for Genealogy – Getting Started, Part Three*. http://www.geni.com/blog/dna-testing-for-genealogy-getting-started-part-three-376261.html : accessed 18 March 2016.

Moore, CeCe. *DNA Testing for Genealogy – Getting Started, Part Two*. http://www.geni.com/blog/dna-testing-for-genealogy-getting-started-part-two-376163.html : accessed 18 March 2016.

Moore, CeCe. *Your Genetic Genealogist*. http://www.yourgeneticgenealogist.com/ : accessed 28 November 2015.

Mount, Steve. *Genetic Genealogy and the Single Segment*. http://ongenetics.blogspot.co.uk/2011/02/genetic-genealogy-and-single-segment.html : accessed 11 May 2016.

National Genealogical Society. *2010 NGS Family History Conference Program*. https://web.archive.org/web/20100425064229/http://members.ngsgenealogy.org/Conferences/2010Program.cfm : accessed 30 July 2016.

Owston, Jim. *The Lineal Arboretum: Phasing the X-Chromosome*. http://linealarboretum.blogspot.com/2012/11/phasing-x-chromosome.html : accessed 27 October 2015.

Oxford Dictionaries. (2016) *Oxford Dictionaries*. http://www.oxforddictionaries.com/ : accessed 21 May 2016.

Park, Danny S, Baran, Yael, Hormozdiari, Farhad, et al. (2015) PIGS: improved estimates of identity-by-descent probabilities by probabilistic IBD graph sampling. *BMC Bioinformatics*. 16 (Suppl 5). p. S9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402697/ : accessed 13 February 2015.

Petrone, Justin. *Ancestry.com Shutters SMGF Database Amid Murder Case Controversy*. https://www.genomeweb.com/applied-markets/ancestrycom-shutters-smgf-database-amid-murder-case-controversy : accessed 19 October 2015.

PGP Global Network. *Personal Genome Project: Harvard*. http://www.personalgenomes.org/harvard : accessed 30 January 2016.

Phillips, Andelka M. (2016) Only a click away — DTC genetics for ancestry, health, love…and more: A view of the business and regulatory landscape. *Applied & Translational Genomics*. 8, Personal Genomics: Complications and Aspirations pp. 16–22. http://www.sciencedirect.com/science/article/pii/S2212066116300011 : accessed 30 March 2016.

Prairielad. (2015) *Pseudo/False Segments under 5cM*. Family Tree DNA Forums > Universal Lineage Testing (Autosomal DNA) > Family Finder Advanced Topics, 3 October 2015 1:50 AM. http://forums.familytreedna.com/showthread.php?t=38586 : accessed 18 February 2016.

Purcell, Shaun, Neale, Benjamin, Todd-Brown, Kathe, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*. 81 (3). pp. 559–575. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/ : accessed 14 February 2016.

Rakow, Paul. (2015) *DNA simulation, version 1*. http://www.isogg.org/w/images/4/4a/DNAsimV1.pdf : accessed 12 March 2016.

Raman, 'Tenali. *Centimorgans or Percentages?* http://hoosierdaddywaitingimpatiently.blogspot.com/2015/01/centimorgans-or-percentages.html : accessed 10 December 2015.

Richard Hill. *DNA Relationship Data*. http://www.dna-testing-adviser.com/DNA-Relationship-Data.html : accessed 6 December 2015.

Roach, Jared C., Glusman, Gustavo, Hubley, Robert, et al. (2011) Chromosomal Haplotypes by Genetic Phasing of Human Families. *The American Journal of Human Genetics*. 89 (3). pp. 382–397. http://www.sciencedirect.com/science/article/pii/S0002929711003181 : accessed 31 October 2015.

Roberson, Elisha D. O. & Pevsner, Jonathan. (2009) Visualization of Shared Genomic Regions and Meiotic Recombination in High-Density SNP Data. *PLoS ONE*. 4 (8). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2725774/ : accessed 14 February 2016.

Rodriguez, Jesse M., Bercovici, Sivan, Huang, Lin, et al. (2015) Parente2: a fast and accurate method for detecting identity by descent. *Genome Research*. 25 (2). pp. 280–289. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315301/ : accessed 15 February 2016.

Rose, Christine. (2009) *Genealogical Proof Standard: Building a Solid Case*. 3rd Ed. Revised. San Jose, California: C R Publications.

Russell, Judy G. *2015: Most bang for the DNA buck*. http://www.legalgenealogist.com/blog/2015/02/02/2015-most-bang-for-the-dna-buck/ : accessed 18 March 2016.

Schaffner, Stephen F. (2004) The X chromosome in population genetics. *Nature Reviews Genetics*. 5 (1). pp. 43–51. http://www.broadinstitute.org/~sfs/nrg_Xchrom.pdf : accessed 17 June 2016.

Southern California Genealogical Society. *Jamboree 2016*. http://www.genealogyjamboree.com/ : accessed 30 January 2016.

Speed, Doug. *Who's your cousin? Using DNA to determine relatedness (Doug Speed).* https://youtu.be/zYAvPuQd0Y8 : accessed 29 April 2016.

Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44. http://dougspeed.com/wdytya/ : accessed 29 April 2016.

Stevens, Eric L., Heckenberg, Greg, Roberson, Elisha D. O., et al. (2011) Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. *PLoS Genetics*. 7 (9). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178600/ : accessed 14 February 2016.

Stone, Dan. *Using GEDmatch (The Triangulation Tool)*. https://stonefamilytree.wordpress.com/2015/01/10/using-gedmatch-the-triangulation-tool/ : accessed 23 November 2015.

Thomas, Thad. (2015) *Re: Thad Thomas - MSc topic*. E-mail to Alasdair Macdonald and Graham Holton. 2:23 AM. 29 August. thad.thomas.2013@uni.strath.ac.uk.

Thompson, Elizabeth A. (2013) Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*. 194 (2). pp. 301–326. http://www.genetics.org/content/194/2/301 : accessed 14 February 2016.

Turner, Ann. *Ahnentafel numbers of ancestors who could contribute a segment on the X* *chromosome.* http://dnacousins.com/AHN_X.TXT : accessed 19 December 2015.

Utah Genealogical Association. *[SLIG] Tracks*. https://web.archive.org/web/20130120204902/http://www.infouga.org/aem.php?lv=p&epg=27 : accessed 30 July 2016.

Vergano, Dan. *This Man Failed A Paternity Test Due To His Vanished Twin's DNA*. http://www.buzzfeed.com/danvergano/failed-paternity-test-vanished-twin : accessed 16 November 2015.

Visscher, P. & Haley, C. (2001) True and false positive peaks in genomewide scans: The long and the short of it. *Genetic Epidemiology*. 20 (4). pp. 409–414. http://dx.doi.org/10.1002/gepi.1010 : accessed 13 February 2016.

Vocabulary.com. *meiosis*. https://www.vocabulary.com/dictionary/meiosis : accessed 13 June 2016.

Waldron, Paddy. *Measuring the length, the rarity and the relevance of shared autosomal DNA*. http://www.pwaldron.info/DNA/significance.html : accessed 6 December 2015.

Walters, Kevin & Cannings, Chris. (2005) The probability density of the total IBD length over a single autosome in unilineal relationships. *Theoretical Population Biology*. 68 (1)., John Maynard Smith Memorial Issue pp. 55–63. http://www.sciencedirect.com/science/article/pii/S0040580905000523 : accessed 14 February 2016.

Walton, Gregory. (2015) Queen intervenes to settle title feud opening way to title pretenders. *Telegraph*. 11 October. http://www.telegraph.co.uk/news/uknews/queen-elizabeth-II/11925101/Queen-intervenes-to-settle-title-feud-opening-way-to-title-pretenders.html : accessed 16 November 2015.

Watterson, Bill. (1996) *There's Treasure Everywhere: A Calvin and Hobbes Collection*. New York: Scholastic Inc.

Wellcome Trust Sanger Institute. *Timeline: History of genomics*. http://www.yourgenome.org/facts/timeline-history-of-genomics : accessed 31 May 2016.

Wikimedia Foundation, Inc. *Assay*. https://en.wikipedia.org/w/index.php?title=Assay&oldid=724116360 : accessed 13 June 2016. Page Version ID: 724116360.

Wikimedia Foundation, Inc. *Centimorgan*. https://en.wikipedia.org/wiki/Centimorgan : accessed 2 December 2015.

Worth, Don. [Response to Roberta Estes blog post]. *DNAeXplained – Genetic Genealogy: Testing Ancestry's Amazing 'New Ancestor' DNA Claim, 4 January, 8:45 PM.* http://dna-explained.com/2015/04/07/testing-ancestrys-amazing-new-ancestor-dna-claim/ : accessed 12 March 2016.

Zhang, Sarah. *Illumina, the Google of Genetic Testing, Has Plans for World Domination*. http://www.wired.com/2016/02/gene-sequencing-goliath-wants-get-bigger-still/ : accessed 3 March 2016.

Zielinski, Dina. *It takes more than a village*. https://blog.dna.land/2016/02/11/it-takes-more-than-a-village/ : accessed 13 February 2016.

Zielinski, Dina. (2016) 'Rule your genome: democratize health.' *RootsTech 2016*. Salt Lake City 6 February 2016. http://www.slideshare.net/DinaZielinski/roots-tech-2016 : accessed 13 February 2016.

APPENDIX A: FINDING AN ADOPTEE'S BIOLOGICAL FAMILY

An adoptee—identified here as an8181—decided to pursue the identity of her biological parents. Her adoptive parents had been provided with a document from the adoption agency that contained information about her birth parents.[268] This information included various genealogical clues that could be helpful in vetting candidate parents including: information about her own birth, her birth parent's ages, marital status at the time of the adoption, the marital status of each of her biological grandparents, physical descriptions of her biological parents and grandparents, occupational information about her biological parents and grandparents.

an8181 submitted a sample for genotyping to AncestryDNA™. When she received her genotyping results at the end of December 2015, her closest match (designated C.S.) was estimated to be a 3rd - 4th cousin—sharing 153 cM across 7 segments.[269] This match had limited access to their published genealogy, and they were unresponsive to requests for access.

On January 23rd, 2016, an8181's match list was revisited. A new, closer match (designated M.G.) had been added to her match list. This match was given as a 1st - 2nd cousin—sharing 454 cM across 22 segments. Using the distributions published from Bettinger's *Shared cM Project* (Figure 10), the 454 cM of shared atDNA falls in the peak of the distribution of Degree 4 relationships.

The administrator of M.G.'s genotype had associated his genotype with a publicly accessible genealogy. Browsing close relatives of M.G., an8181's husband was able to find a candidate family with a son that could have been a sibling to one of an8181's parents. The administrator of M.G.'s genotype was contacted, and the administrator was able to confirm that M.G.'s half first cousin had given up a child for adoption on the day an8181 was born. M.G.'s half first cousin was an8181's biological father.

Genealogical research about an8181's biological father was able to reveal an8181's mother.[270] Many of the details given in the document from the adoption agency about an8181's biological parents could be identified in the genealogical records found about her parents and their families.

an8181's search for her biological parents turned out to be a relatively straight forward process. It is typical for an adoptee to have to wait for the right match, but often the wait is much longer than occurred here. There are strategies for actively seeking better matches, but she did not have to employ any of these. Additional triangulation is often required, but the nature of the match combined with the details available from the adoption agency and from members of the biological family rendered this level of proof unnecessary.

For more information on methodology around seeking biological parents as an adoptee, the genetic genealogical community recommends starting with the resources published at DNAadoption.com.[271,272,273]

APPENDIX B: TRIANGULATION FOR GT999 ON CHR1 FROM 159M TO 167M

The following triangulated group was identified with the phased maternal kit for GT999 on Chr1:

KIT #	START POS	END POS	cM	SNPs
GT611	72,017	247,169,190	281.5	54,743
GT861	99,175,785	194,953,541	73.8	16,350
GT709	111,732,674	194,928,236	60.8	13,486
GT610	114,435,751	200,890,245	64.9	13,883
GT901	152,790,212	194,928,236	45.5	9,809
GT124	155,348,641	166,650,792	20.3	3,348
GT732	158,328,558	166,650,792	14.8	2,522
GT831	158,879,805	177,283,021	22.3	4,848

Figure 34: Triangulated group with GT999 on Chr1 for a matching segment from 159M to 167M.[274]

All the individuals in the group share a common ancestor as shown in Figure 35.

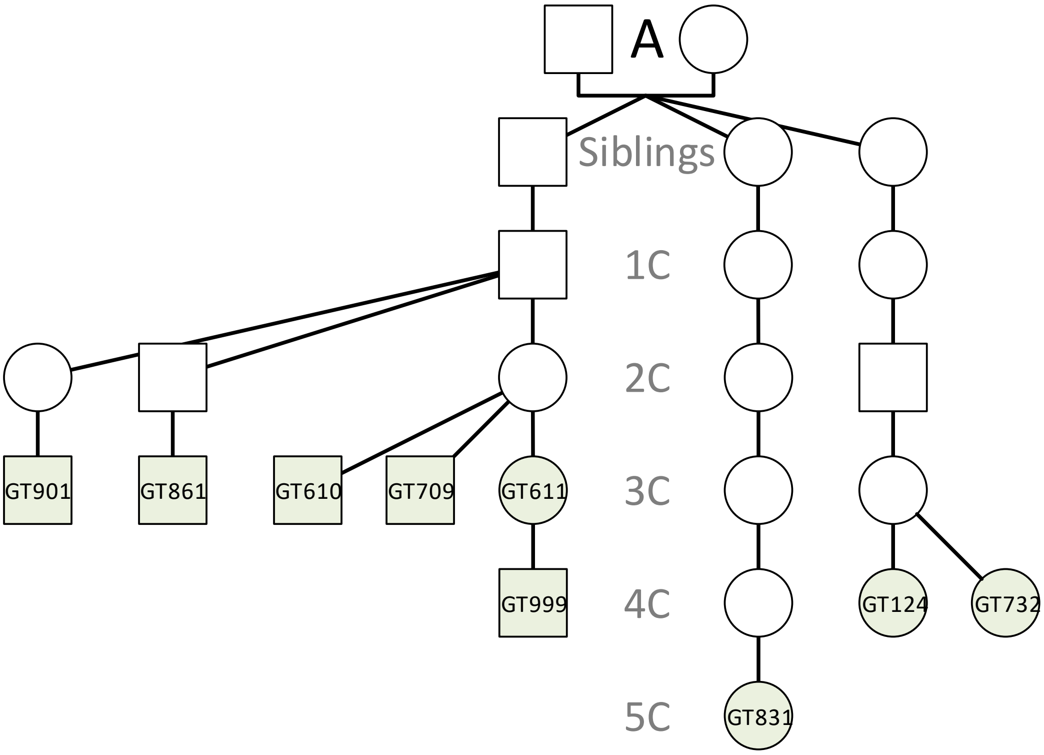


Figure 35: Representation of the lineages connecting members of a triangulated group for Chr1 (159M to 167M).[275,276,277] *See also Figure 13, Figure 14, and Figure 18.*

This triangulated group has a large family group (to be called Group 1 in this appendix) composed of GT999, GT611, GT709, GT610, GT861, and GT901. There is another family group (to be called Group 2 in this appendix) composed of two sisters: GT124 and GT732. A third individual GT831 is also a member of the triangulated group.

There is another individual—GT681—who matches the group, but whose common ancestor with the group has not been identified. Therefore, GT681 has not been represented in Figure 34 or Figure 35; GT681 is present in Figure 37. Several candidate common ancestors with GT999 have been identified that are ninth great-grandparent generation for GT999, suggesting that GT681 has a much more distant relationship with members of this group. This group may be candidate for an intermediate MRCA if another match is found that will triangulate with GT681.

Using the matches listed in Figure 34, the segment shared by this group is positioned from 158,879,805 (159M) to 166,650,792 (167M) with a physical length of 7,770,987 bps (7.8 Mbp) and a genetic size of 13.2096 cM.[278]

The *hypothesis* is that Group 1, Group 2 and GT831 all received the shared segment IBD from the common ancestor identified for the triangulated group. Before accepting the *hypothesis*, factors affecting the likelihood of reliability, errors or misinterpretations must be considered; also, the items of *evidence* should agree with each other and with the principles of inheritance. The testing that follows will help determine whether the *hypothesis* can be accepted as a *conclusion*.

TESTS OF ANALYSIS

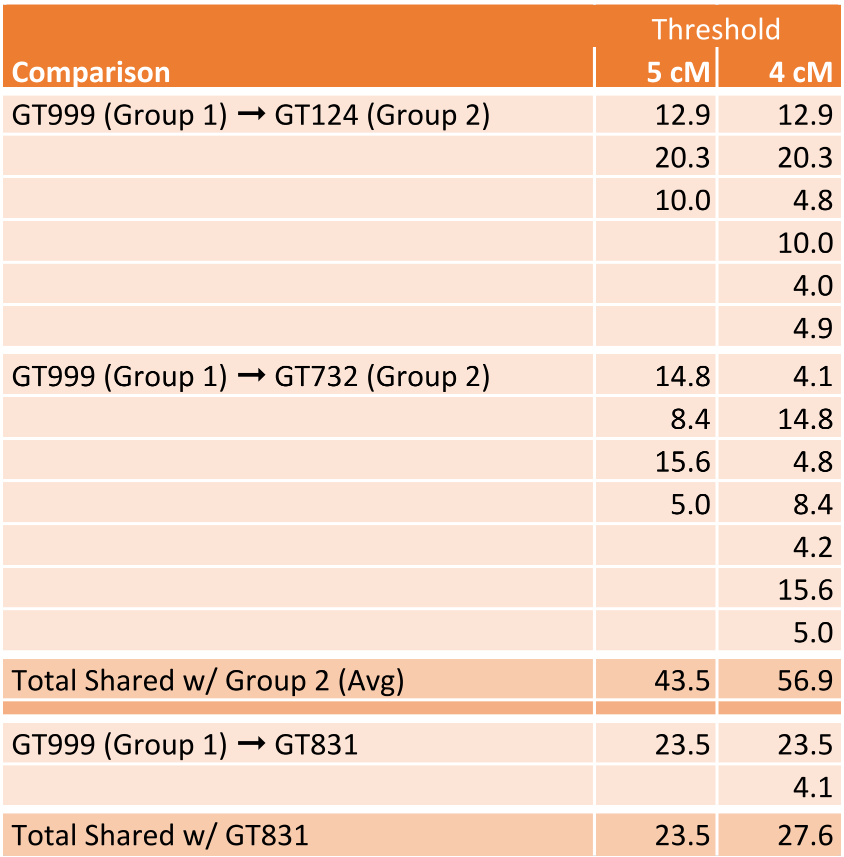
MATCHING SEGMENT SIZE

The size of the individual matches shared with GT999 are all greater than 20 cM except for GT732 at 14.8 cM. Based on Figure 19, the probability any of these matches will not survive phasing is negligible. As the comparisons were all made with GT999's phased maternal genotype, it is very unlikely that these matches are IBC. All are well above the minimum 5.0 cM threshold suggested for phased matches, the threshold necessary to achieve an acceptable level of false positives.[279] The matches are certainly all good candidates to be IBD according to these tests.

The physical length of the matches that GT999 shares with Group 2 and GT831 range from 8,322,234 bps (8.3 Mbp) to 18,403,216 bps (18.4 Mbp). Considering these matches in the context of Speed and Balding's work (Figure 23), both relationship types are considered to have *G*=5. The smallest match is in the tranche from 5-10 Mbp, and the largest is in the tranche from 10-20 Mbp. For 10-20 Mbp tranche, there is more than a 10% chance that a match of the given size came from an ancestor with *G*=5 or closer. For the 5-10 Mbp tranche, the likelihood has fallen to roughly 5%. In all cases, the probability is not negligible; the matches are candidates to be IBD.

TOTAL SHARED IBD

The theoretical average sharing for fourth cousins (the relationship between GT999 in Group 1, and Group 2) is 13.28 cM.[280] For fourth cousins once removed (GT999 in Group 1 and GT831), it is 6.64 cM. In all cases, the match from each individual that is being evaluated in this triangulated group already exceed these values. This comparison is not meaningful.



	Threshold	
Comparison	**5 cM**	**4 cM**
GT999 (Group 1) ➞ GT124 (Group 2)	12.9	12.9
	20.3	20.3
	10.0	4.8
		10.0
		4.0
		4.9
GT999 (Group 1) ➞ GT732 (Group 2)	14.8	4.1
	8.4	14.8
	15.6	4.8
	5.0	8.4
		4.2
		15.6
		5.0
Total Shared w/ Group 2 (Avg)	43.5	56.9
GT999 (Group 1) ➞ GT831	23.5	23.5
		4.1
Total Shared w/ GT831	23.5	27.6

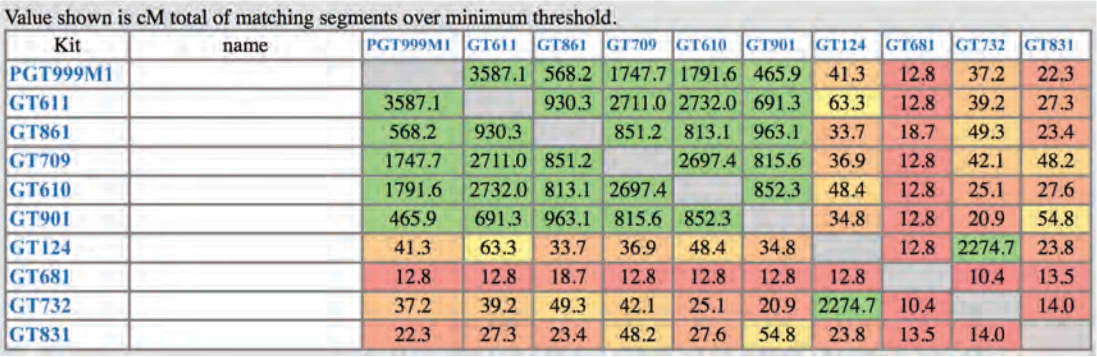
Figure 36: GT999's total atDNA sharing with Group 2 (an average) and with GT831.[281]

GT999's total amount of sharing with members of Group 2 (calculated as an average per individual in the group) and with GT831 (given in Figure 36) was calculated using Janzen's method.[282] Both relationships are outside the relationships reported in Figure 21 (from the *Shared cM Project*). The shape and bounds of the distributions from the *Shared cM Project* for these relationship types are difficult to discern in Figure 10, but the associated distributions seem to cover the calculated totals for both relationships being considered here. The totals are high, presumably because they include some amount of IBC (false) sharing or because the relatives share more atDNA that would be expected by chance. The difference between the amounts shared matches the expected difference—the fourth cousins sharing roughly twice as much atDNA than the fourth cousin once removed.

The total amount of sharing is consistent with the relationships specified in the *hypothesis*.

TESTS OF CORRELATION

INDEPENDENCE



Value shown is cM total of matching segments over minimum threshold.

Kit	name	PGT999M1	GT611	GT861	GT709	GT610	GT901	GT124	GT681	GT732	GT831
PGT999M1			3587.1	568.2	1747.7	1791.6	465.9	41.3	12.8	37.2	22.3
GT611		3587.1		930.3	2711.0	2732.0	691.3	63.3	12.8	39.2	27.3
GT861		568.2	930.3		851.2	813.1	963.1	33.7	18.7	49.3	23.4
GT709		1747.7	2711.0	851.2		2697.4	815.6	36.9	12.8	42.1	48.2
GT610		1791.6	2732.0	813.1	2697.4		852.3	48.4	12.8	25.1	27.6
GT901		465.9	691.3	963.1	815.6	852.3		34.8	12.8	20.9	54.8
GT124		41.3	63.3	33.7	36.9	48.4	34.8		12.8	2274.7	23.8
GT681		12.8	12.8	18.7	12.8	12.8	12.8	12.8		10.4	13.5
GT732		37.2	39.2	49.3	42.1	25.1	20.9	2274.7	10.4		14.0
GT831		22.3	27.3	23.4	48.2	27.6	54.8	23.8	13.5	14.0	

Figure 37: Total amount of atDNA sharing between members of the group that triangulates on a matching Chr1 segment from 159M to 167M.[283]

The report shown in Figure 37 helps evaluate the independence within the triangulated group. There are two family groups (Group 1 includes the first six individuals; Group 2 includes the seventh and ninth individuals), and two individuals (GT831 and GT732). Only three independent answers to the triangulation question are possible. So far, only two answers have been identified.

A common ancestor has not been identified for GT681. Note that GT681's matches with the group are generally smaller (her row and column in the table is a much darker red). As already mentioned, the candidate common ancestors that have been identified thus far for GT681 are much more distant than the ancestor identified for this group. The smaller size of her matches with the group seems to be indicative of this more distant relationship.

Excessive Matching

With only three possible instances of $\mathcal{E}$ in the database thus far, excessive matching is not a worry for this triangulated group.

Chromosome Map Correlation

Figure 31 is a Chr1 map for four siblings, three of which are members of this triangulated group: GT611, GT709 and GT610. The segment under consideration lies somewhere to the left of the yellow line marking 175M. The next numeric reference point along the map marks 158M which corresponds to the crossover from green to brown on the person labeled *P*. The segment in question lies somewhere between 158M and 175M, but on the regions colored light blue. There are no crossover events within this range. Within this range, there is only one possible configuration of siblings that could match the light-blue region—and the expected three are the ones with matching segments. The chromosome map agrees with the IBD claim.

Intermediate Common Ancestors

GT999 has intermediate common ancestors that share this segment at his grandmother (two uncles: GT709 and GT610), and at his great-grandfather (two of his mother's cousins: GT861 and GT901). The intermediate common ancestors are strong *evidence* that this segment was received IBD.

Phased Matching

The phased maternal genotype of GT999 was used to identify members of this triangulated group, essentially eliminating the risk that IBC matches were included in the group—a boost to the IBD claim.

Generational Matching

Both GT999 and his mother GT611 match with Group 2 and with GT831. The expected generational match is present—*evidence* that the matching segment was received IBD.

GT999 also has a son GT186, but the son does not match any of the members of Group 2, or GT831.[284] This does not cast doubt on our hypothesis because it is plausible that the son did not receive this segment.

Close Relative Matching

GT999 has several close relatives that match the individuals in Group 2, and that match GT831: two uncles (GT709 and GT610), and two of his mother's cousins (GT861 and GT901). These matches strengthen the claim that GT999 received this segment IBD.

The individuals in Group 2 (GT124 and GT732) are siblings and they both match this shared segment—*evidence* that they received this segment IBD.

Match Stability

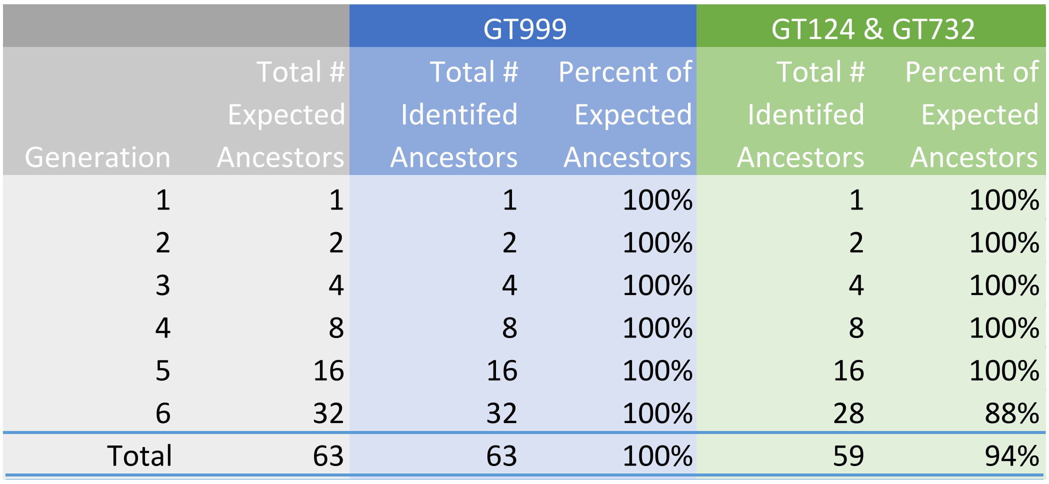
Match stability can only be considered in the context of GT999 and his mother GT611. Comparing mother and son to Group 2 and to GT831, the match is stable—genetically correct; i.e., the newest generation received a segment that was the same size as or smaller than the older generation, and the newer generation's segment was bounded at or within the boundaries of the older generation—as it was passed from mother (GT611) to son (GT999).[285,286]

A lack of stability would have cast doubt on an IBD claim. There is no lack of stability in this case.

Ɛ (1 of 2)

This instance of Ɛ is between Group 1 and Group 2.

COMMON ANCESTOR UNIQUENESS



		GT999		GT124 & GT732	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	28	88%
Total	63	63	100%	59	94%

Figure 38: Compiled genealogy completeness evaluation for GT999 and Group 2 (GT124 & GT732).[287,288]

The compiled genealogy representing Group 1 is complete to the generation of the proposed common ancestor. The compiled genealogy representing Group 2 is only missing the identities of four ancestors, all of them missing in the last generation. There is little risk that a different MRCA will be identified between these two family groups.

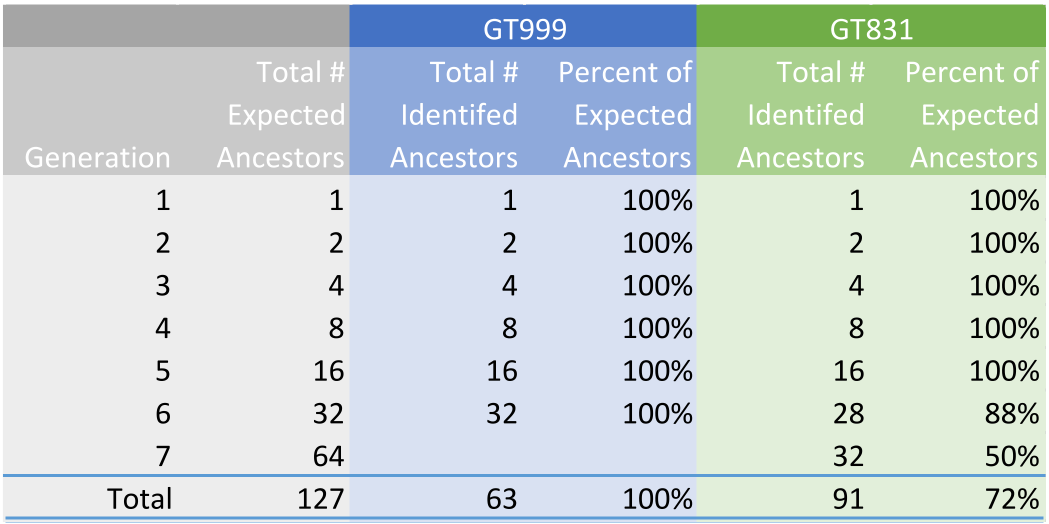
CONCLUSION FOR ε (1 OF 2)

All the testing for match (m) strongly supports that this shared segment is consistent with IBD segments, and that it was received IBD. The search for a common ancestor (a) between these two groups has considered all but four ancestors from Group 2—a very small residual risk.

ε (2 OF 2)

This instance of ε is between Group 1 and GT831.

COMMON ANCESTOR UNIQUENESS



		GT999		GT831	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	28	88%
7	64			32	50%
Total	127	63	100%	91	72%

Figure 39: Compiled genealogy completeness evaluation for GT999 and GT831.

The compiled genealogy representing Group 1 is complete to the generation of the proposed common ancestor. The compiled genealogy representing GT831 is missing half of the identities of in the last generation and four identities in the preceding generation. There are definitely gaps in this search.

CONCLUSION FOR ε (2 OF 2)

All of the testing for match (m) strongly supports that this shared segment is consistent with IBD segments, and that it was received IBD. The search for a common ancestor (a) between these two groups has gaps. Some mitigation is had in the fact that there are two instances of $\mathcal{E}$ in this hypothesis (each with independent instances of m and independent lineages to a), which is rare in and of itself; this tends to offset the risk that lingers due to the lack of coverage in the MRCA search.

Conclusion Accepting

There is every reason to believe that the specified segment was received IBD by members of this triangulated group. The proposed common ancestor is unique with little residual risk for one instance of $\mathcal{E}$, but the genealogy of GT831 is incomplete and exposes the *hypothesis* to additional risk. This risk is mitigated by the fact that there are two independent instances of $\mathcal{E}$ in the hypothesis. It is therefore *concluded* that one person in the ancestral couple identified as common to this triangulated group must have contributed this shared segment (Chr1 from 159M to 167M) to all of the members of this triangulated group.

APPENDIX C: TRIANGULATION FOR GT999 ON CHR4 FROM 187M TO 191M

The following triangulated group was identified with the phased paternal genotype for GT999 on Chr4:

KIT #	START POS	END POS	cM	SNPs
GT163	61,566	191,117,403	214.4	37,262
GT625	175,696,978	190,696,128	36.0	4,166
GT177	175,740,345	190,696,128	35.9	4,128
GT381	184,520,405	190,922,297	19.1	1,819
GT136	185,592,550	190,568,137	14.9	1,426
GT978	186,651,620	191,117,403	12.0	1,207
GT789	186,651,620	191,117,403	12.0	1,207
GT606	186,651,620	191,117,403	12.0	1,212
GT491	186,651,620	191,117,403	12.0	1,199

Figure 40: Triangulated group with GT999 on Chr4 for a matching segment from 187M to 191M.[289]

With the exception of GT136 (whose connection with the group is currently undetermined), all share a common ancestor as shown in Figure 27.

The match is positioned from 186,651,620 (187M) to 190,568,137 (191M). It has a physical length of 3,916,517 bps (3.9 Mbp) and a genetic size of 11.0543 cM.[290]

TESTS OF ANALYSIS

MATCHING SEGMENT SIZE

Based on Figure 19, the probability that any of the matches is IBC is less than 5%. As the comparisons were made with a phased genotype, it is already known that these matches are not IBC. All are well above the minimum 5.0 cM threshold suggested for phased matches, the threshold necessary to achieve an acceptable level of false positives.[291] The matches are certainly good candidates to be IBD.

TOTAL SHARED IBD

For all of the genotypes in the group (except GT163—father of GT999), there is only one shared segment with GT999. A comparison with a theoretical average for total sharing is rendered meaningless because the closest relationship with GT999 (that is not his father) is expected to share just 3.32 cM which is below the segment reliability thresholds already mentioned, and it is outside the relationships reported by the Shared cM Project.[292,293]

Instead, whether the match size is IBD-consistent is considered within context of Speed and Balding's work.[294] A 3.9 Mbp shared region falls in the tranche between 2 and 5 Mbp (Figure 23)—a small, shared region by IBD-consistent standards. Roughly 10% of shared segments this size belong to ancestors in the closest 10 generations. While it is considered rare to share such a small segment IBD from a G=9 ancestor, it is not an impossible occurrence. The shared segment's genetic size (11.1 cM) suggests that an IBD segment in this region will generally be quite small physically.

Tests of Correlation

Independence

The report shown in Figure 26 helps evaluate independence within the group.[295] It shows that there are three family groups and one individual. GT999 is in a group with his father (GT163). There can be three independent answers to the triangulation question between GT999 and the other independent groups/individuals.

Excessive Matching

Based on in Figure 26, the number of independent matches at this location is only three; this does not seem to be an excessive amount of matching.

Chromosome Map Correlation

GT999 does not have an established chromosome map for any other generations on this chromosome, so no correlation with a chromosome map is possible.

Intermediate Common Ancestors

At this time, there are no known intermediate common ancestors that correlate with this matching segment.

Phased Matching

The matching was done with the phased paternal genotype of the POI, essentially eliminating the risk that IBC matches were included in the group—giving a boost to the IBD claim.

$\mathcal{E}$ (1 of 3)

The first instance of $\mathcal{E}$ is between the group headed by GT163 (with son GT999) and the group headed by GT978 (including child GT789, and grandchildren GT606 and GT491). All of the members of this group match each other in the expected ways (meaning all of the parents share the expected amount with their children, and the siblings share an expected amount with their sibling).

GENERATIONAL MATCHING

Generationally, the phased matching already guarantees a generational match between GT999 and GT163. GT999 does have a son (GT186) who could have been included in the group; his phased genotype matches the GT978-group in the same stable manner as GT999's genotype.[296] There is no additional generational matching that can be tried between these two family groups. This testing supports the IBD claim.

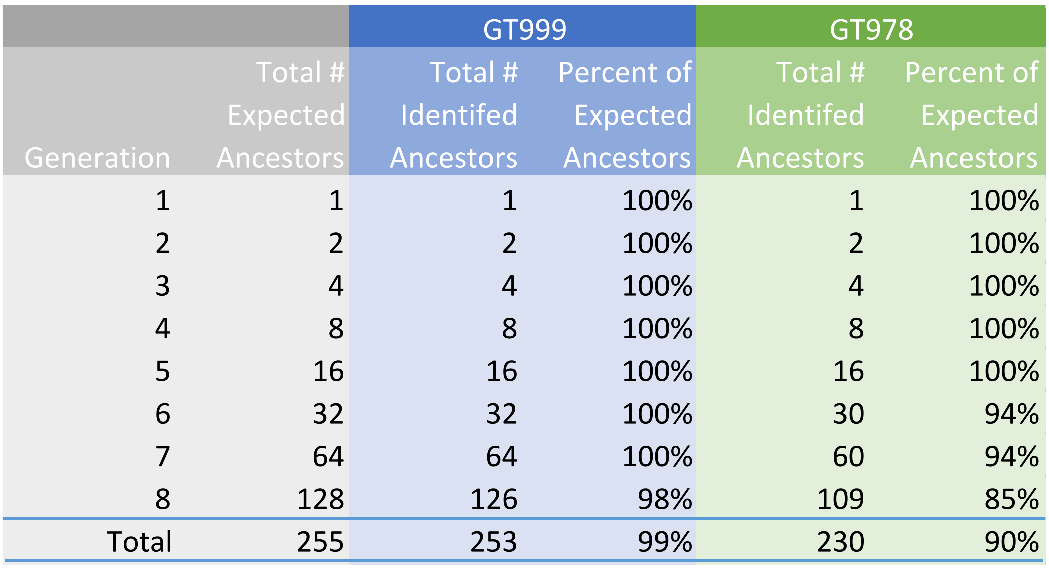
CLOSE RELATIVE MATCHING

This group did not match any genotypes of other identified close relatives for GT999 (i.e., an aunt, a second cousin and twice-related cousin—3C /3C1R).[297] A match could have strengthened the IBD claim; a lack of match does not cast any appreciable doubt.

MATCH STABILITY

The matches of the phased genotype of GT999 with the genotypes of the GT978-group are very stable—all of them starting and ending at the same position—giving credibility to the *hypothesis* that these matches are IBD-consistent.

COMMON ANCESTOR UNIQUENESS



		GT999		GT978	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	30	94%
7	64	64	100%	60	94%
8	128	126	98%	109	85%
Total	255	253	99%	230	90%

Figure 41: Compiled genealogy completeness evaluation for GT999 and GT978.[298,299]

For eight generations, GT999 and GT978 both have great coverage in their compiled genealogies—GT999 has identified 99% of his ancestors, and GT978 has identified 90% of his ancestors. GT978 does not reference much in the way of original records, but other compiled genealogies for his lineage to the common ancestor do.[300] No other factors were identified that might affect the reliability of the information contributing to the lineages that connect these two cousins. Some risk remains but would be considered limited relative to many other cases.

CONCLUSION FOR $\mathcal{E}$ (1 OF 3)

All the tests of analysis and correlation for the match (m) for this first instance of $\mathcal{E}$ support the IBD claim—none presenting any conflicts, and only the map correlation test being untried. The biggest risk for the match (m) seems to be the physical length of the matching segment—it being relatively smaller than perhaps expected. The risks associated with the compiled genealogies are also relatively small. The first instance of $\mathcal{E}$ seems to bear up well to scrutiny.

$\mathcal{E}$ (2 OF 3)

The second instance of $\mathcal{E}$ is between the group containing GT999 and GT1653 and the group containing GT381, GT625 and GT177. GT381 and GT625 are half-siblings, and GT177 is their half-aunt.[301] GT381 matches her half-sibling and half-aunt at the high-end of their expected total sharing amounts—to be expected given there is no phasing to eliminate IBC matching—supporting the relationships claimed within this family group.[302]

GENERATIONAL MATCHING

The phased matching already guarantees a generational match GT999 and GT163. GT999's son also matches the members of this family group.[303] There is no additional generational matching that can be tried between these two family groups. This testing supports the IBD claim.

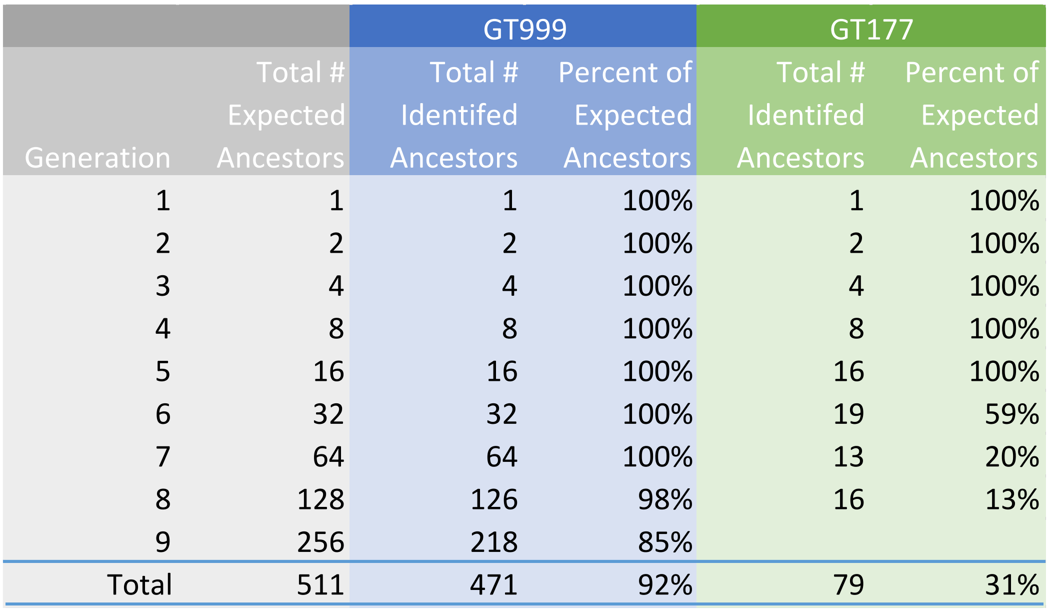
CLOSE RELATIVE MATCHING

GT381 and GT625 did not match any of the genotypes of other close relatives (an aunt, a second cousin and twice-related cousin—3C /3C1R), but GT625 does match the second cousin (GT859) nearby (Chr4 from 176M to 184M).[304] This match strengthens the claim of relatedness, and in so doing, strengthens the claim that the segment under consideration was received IBD.

MATCH STABILITY

The test cannot be applied because the members of the group being tested are not lineally related.

COMMON ANCESTOR UNIQUENESS



		GT999		GT177	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	19	59%
7	64	64	100%	13	20%
8	128	126	98%	16	13%
9	256	218	85%		
Total	511	471	92%	79	31%

Figure 42: Compiled genealogy completeness evaluation for GT999 and GT177.[305,306]

While GT999's genealogy is fairly complete, GT177's genealogy is very incomplete. This does tend to cast doubt on the completeness of the MCRA search. Though incomplete, the genealogy covering GT177's lineage to the proposed common ancestor does appear to be original research and does reference expected original records. No other factors were identified that might affect the reliability of the information contributing to the lineages that connect these two cousins.

CONCLUSION FOR $\mathcal{E}$ (2 OF 3)

The second instance of $\mathcal{E}$'s main flaw is the coverage of ancestors in the search for the MRCA. The fact that there are two instances of $\mathcal{E}$ is rare in and of itself; this tends to offset the risk that lingers due to the lack of coverage in the MRCA search. The second instance of $\mathcal{E}$ bears up to scrutiny, but there is a bit more risk that future research may introduce conflict.

$\mathcal{E}$ (3 OF 3)

COMMON ANCESTOR UNIQUENESS

The third instance of $\mathcal{E}$ is between the group containing the group headed by GT163 and GT136. Attempts to identify a common ancestor between GT136 and other members of this triangulated group have failed. The administrator of this genotype has only been able to share the genealogy of the paternal grandfather of GT136, leaving 75% of the ancestors of GT136 unidentified. Furthermore, the administrator is aware of four other genotypes that are related to GT136 and that all match GT136 and that are related to this paternal grandfather.[307] None of the members of this triangulated group match any of the other four genotypes—making it much less likely that members of this triangulated group have a relationship with this paternal grandfather.[308] The common ancestor is most likely among the unidentified ancestors of GT136.

GENERATIONAL MATCHING

GT136's match with the GT999 is not IBC because of the comparison with the GT999's phased genotype. This also guarantees a generational match with the GT999's father. The GT999's son is also matches with GT136 on this shared segment.[309] No additional generational matching is possible at this time.

CLOSE RELATIVE MATCHING

GT136 did not match any other close relatives.[310]

CONCLUSION FOR $\mathcal{E}$ (3 OF 3)

GT136 remains a member of the triangulated group, and the match should be considered when evaluating excessive matching; but it could end up in conflict with the group, or as an intermediate MRCA, or even as an instance of $\mathcal{E}$ for an older generation.

OTHER CONSIDERATIONS

One additional risk must be called out. The common ancestor in the first instance of $\mathcal{E}$ is one generation newer that the common ancestor in the second instance of $\mathcal{E}$. This is not ideal, leaving room for doubt. It could be argued that the first instance of $\mathcal{E}$ should be represented as an intermediate MRCA for the second instance of $\mathcal{E}$—leaving only one instance of $\mathcal{E}$ in the triangulated group. But lacking other matches, and lacking any conflict, and given the relative independence of these two instances, it seems safe to accept both instances of $\mathcal{E}$ for the purposes of triangulating this shared segment until such time as a proper instance of $\mathcal{E}$ presents itself (at which time the first instance can become an intermediate MRCA), or until a conflict overturns the proposed solution.

CONCLUSION ACCEPTING

Given that two of the three instances of $\mathcal{E}$ in this triangulated group pass testing with minimal residual risk, and given that the third instance of $\mathcal{E}$ (without an identified common ancestor with the group) remains a candidate member of the group (insofar as it could be tested), it is therefore *concluded* that one person in the ancestral couple identified as common to this triangulated group must have contributed the Chr4 matching segment (187M to 191M) to the members of this triangulated group.

APPENDIX D: TRIANGULATION FOR GT611 ON CHR4 FROM 177M TO 191M

Figure 43 shows a set of genotypes that match GT611 on Chr1.

KIT #	CA	START POS	END POS	cM	SNPs
GT610	Y	113,623,623	207,063,451	75.0	15,948
GT116		162,639,701	191,953,018	26.9	2,771
GT341		162,926,661	191,859,017	26.2	2,688
GT654	Y	164,598,102	190,827,077	23.2	5,739
GT383	Y	175,935,925	190,828,095	11.3	1,283
GT480	Y	177,088,988	191,031,443	10.8	2,703
GT938		177,936,529	193,743,320	12.9	3,024
GT196		177,936,529	190,681,611	10.0	2,433
GT557		177,936,529	190,681,611	10.0	2,458
GT554		177,945,266	190,677,903	10.0	1,005
GT338		177,936,529	189,944,474	9.6	2,369
GT369		177,578,725	189,891,367	9.7	2,415
GT728		177,936,529	189,891,367	9.6	2,347
GT167		177,936,529	189,891,367	9.6	2,338
GT820		177,936,529	189,002,080	9.4	2,198
GT633		179,386,458	190,680,905	8.8	2,118
GT966		180,689,290	192,059,107	8.5	781

Figure 43: Group of Chr1 matches for GT611 that also matched GT654.[311]

All of the genotypes used in these comparisons are not phased. To ensure all of the individuals match an identical segment (see *Correlating m* on page 60), all of the individuals in the group were also shown to match GT654.

The author has not been able to contact most of the individuals in this list; therefore, for most individuals, a common ancestor remains unidentified. There are few in the group (individuals with a “Y” in the *CA* column in Figure 43) that share a common ancestor as shown in Figure 44. Because all the genotypes in the group passed the correlating comparison, it is unlikely that any are IDC. If all the shared segments turned out to be IBD, we would expect all of the members of this group to have a common ancestor, though the common ancestor may be an ancestor of the common ancestor shared by the four with a known common ancestor.

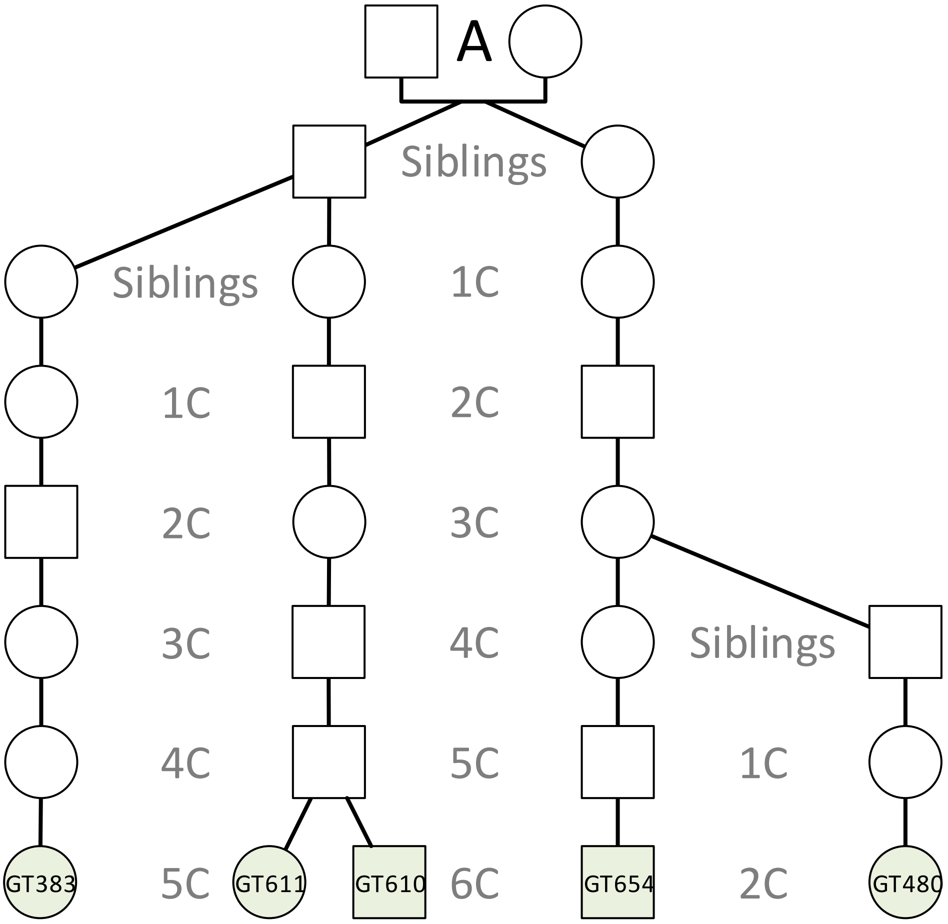


Figure 44: Representation of lineages linking connecting individuals in a triangulated group with GT611 for Chr1 (177M to 191M).[312,313,314,315] *See also Figure 28.*

GT611 and GT610 are siblings and atDNA confirms this relationship. In many instances, GT611 and GT610 can be used interchangeably in the testing detailed in this appendix. Unless GT610 is mentioned explicitly, GT611 can be considered a proxy for GT610 in that test as either sibling's result would be equivalent.

GT654 and GT480 are second cousins. They share 194.1 cM of atDNA.[316] This is consistent with Degree 5 relationship (which includes second cousins) as given in Figure 10.

Using the matches shown in Figure 44, the matching segment for this group is positioned from 177,088,988 (177M) to 190,827,077 (191M) with a physical length of 13,738,089 bps (13.7 Mbp) and a genetic size of 10.5218 cM.[317]

The *hypothesis* is that GT611, GT610, GT654, GT383 and GT480 all received their matching segment IBD from the common ancestor identified for the triangulated group. Before accepting the *hypothesis*, factors affecting the likelihood of reliability, errors or misinterpretations must be considered; also, the items of *evidence* should agree with each other and with the principles of inheritance. The testing that follows will help determine whether the *hypothesis* can be accepted.

Tests of Analysis

Matching Segment Size

The size of the GT611's match with GT654 is greater than 20 cM and (based on Figure 19) the likelihood it will not survive phasing is negligible. The size of the matches with GT383 and GT480 are much smaller and there is roughly a 5-10% chance that these matches would fail to match if compared to a phased genotype. These matches remain candidates to be IBD, but it would be desirable to have other testing to strengthen this claim.

Using physical length, the segment that GT611 shares with GT654 is 26,228,975 bps (26.2 Mbp), and with GT480 is 13,942,455 bps (13.9 Mbp). GT654 and GT480 are sixth cousins with GT611—*G*=7 in the context of Speed and Balding's work (see Figure 22). Using Figure 23, the smallest shared segment is in the tranche from 10-20 Mbp, and the largest is in the tranche from 20-30 Mbp. For the 10-20 Mbp tranche, there is more than a 20% chance that a shared segment of the given size came from an ancestor with *G*=7 or closer. For the 20-30 Mbp tranche, the likelihood is roughly 45%. It is not unreasonable for shared segments of this size to be IBD.

GT383's shared segment has a length of 14,892,170 bps (14.9 Mbp), but a value of *G*=6. Using Figure 23, there is just under a 20% chance that a shared segment of this size is from an ancestor with G=6 or closer. It is not unreasonable for a shared segment of this size to be IBD.

In all cases, the matches remain likely candidates to be IBD.

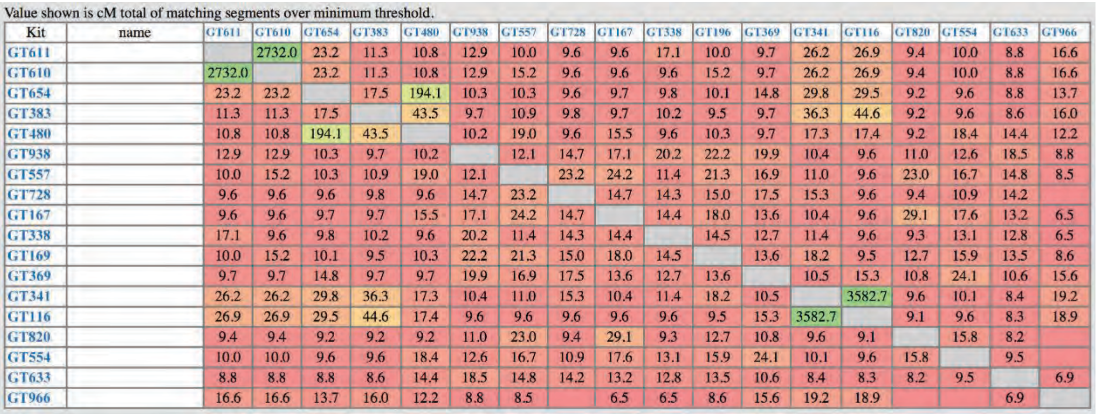
Total Shared IBD

A comparison with a theoretical average for total sharing is rendered meaningless because the closest relationship (GT611 and GT383 are fifth cousins) is expected to share just 3.32 cM which is well below the segment reliability thresholds.[318] The fifth cousin relationship is outside the set of relationships reported by the *Shared cM Project* in Figure 21. The distribution for fifth cousin relationships is difficult to discern in Figure 10, so it is not particularly useful for this situation.

Instead, total IBD sharing testing degrades to an evaluation of the shared segment within the context of Speed and Balding's work.[319] This has already been considered as part of the matching segment size testing and did not cast any doubt on the IBD claim.

Tests of Correlation

Independence



Value shown is cM total of matching segments over minimum threshold.

Kit	name	GT611	GT610	GT654	GT383	GT480	GT938	GT557	GT728	GT167	GT338	GT196	GT369	GT341	GT116	GT820	GT554	GT633	GT966
GT611			2732.0	23.2	11.3	10.8	12.9	10.0	9.6	9.6	17.1	10.0	9.7	26.2	26.9	9.4	10.0	8.8	16.6
GT610		2732.0		23.2	11.3	10.8	12.9	15.2	9.6	9.6	9.6	15.2	9.7	26.2	26.9	9.4	10.0	8.8	16.6
GT654		23.2	23.2		17.5	194.1	10.3	10.3	9.6	9.7	9.8	10.1	14.8	29.8	29.5	9.2	9.6	8.8	13.7
GT383		11.3	11.3	17.5		43.5	9.7	10.9	9.8	9.7	10.2	9.5	9.7	36.3	44.6	9.2	9.6	8.6	16.0
GT480		10.8	10.8	194.1	43.5		10.2	19.0	9.6	15.5	9.6	10.3	9.7	17.3	17.4	9.2	18.4	14.4	12.2
GT938		12.9	12.9	10.3	9.7	10.2		12.1	14.7	17.1	20.2	22.2	19.9	10.4	9.6	11.0	12.6	18.5	8.8
GT557		10.0	15.2	10.3	10.9	19.0	12.1		23.2	24.2	11.4	21.3	16.9	11.0	9.6	23.0	16.7	14.8	8.5
GT728		9.6	9.6	9.6	9.8	9.6	14.7	23.2		14.7	14.3	15.0	17.5	15.3	9.6	9.4	10.9	14.2	
GT167		9.6	9.6	9.7	9.7	15.5	17.1	24.2	14.7		14.4	18.0	13.6	10.4	9.6	29.1	17.6	13.2	6.5
GT338		17.1	9.6	9.8	10.2	9.6	20.2	11.4	14.3	14.4		14.5	12.7	11.4	9.6	9.3	13.1	12.8	6.5
GT169		10.0	15.2	10.1	9.5	10.3	22.2	21.3	15.0	18.0	14.5		13.6	18.2	9.5	12.7	15.9	13.5	8.6
GT369		9.7	9.7	14.8	9.7	9.7	19.9	16.9	17.5	13.6	12.7	13.6		10.5	15.3	10.8	24.1	10.6	15.6
GT341		26.2	26.2	29.8	36.3	17.3	10.4	11.0	15.3	10.4	11.4	18.2	10.5		3582.7	9.6	10.1	8.4	19.2
GT116		26.9	26.9	29.5	44.6	17.4	9.6	9.6	9.6	9.6	9.6	9.5	15.3	3582.7		9.1	9.6	8.3	18.9
GT820		9.4	9.4	9.2	9.2	9.2	11.0	23.0	9.4	29.1	9.3	12.7	10.8	9.6	9.1		15.8	8.2	
GT554		10.0	10.0	9.6	9.6	18.4	12.6	16.7	10.9	17.6	13.1	15.9	24.1	10.1	9.6	15.8		9.5	
GT633		8.8	8.8	8.8	8.6	14.4	18.5	14.8	14.2	13.2	12.8	13.5	10.6	8.4	8.3	8.2	9.5		6.9
GT966		16.6	16.6	13.7	16.0	12.2	8.8	8.5		6.5	6.5	8.6	15.6	19.2	18.9			6.9	

Figure 45: Total amount of atDNA sharing between members of the group that triangulates on a matching Chr1 segment from 177M to 191M.[320]

There are seventeen genotypes that match the segment being examined here. Of this seventeen, there are three pairs of closely related individuals: GT611 and GT610 (siblings), GT654 and GT480 (second cousins), and GT116 and GT341 (a parent/child pair). The remaining individuals appear to be more distantly related. This means that thirteen independent answers to the triangulation question (thirteen independent instances of $\mathcal{E}$) are possible with this set of matches. The current hypothesis has only two of the possible thirteen instances of $\mathcal{E}$.

Excessive Matching

There are fourteen independent matches with GT611 on this segment—which starts to feel like a large number. This may be indicative of excessive matching, which means some of these matches may be IBS. Many matching segments are on the small side, making them more likely to be IBS, or at least making the associated individuals more distantly related (and therefore hard to identify as relations because genealogies do not generally reach back far enough in time to make the necessary connections).

CHROMOSOME MAP CORRELATION

Figure 31 is a Chr1 map for two of the members of the triangulated group: GT611 and GT610. The matching segment under consideration lies somewhere in between the yellow line marking 175M and 195M. The matching segment belongs to the brown-colored regions. There are no crossover events within this range. Within this range, there is only one possible configuration of siblings that could match the brown-colored regions—and the expected two siblings are the ones matching this segment. The chromosome map supports the IBD claim.

INTERMEDIATE COMMON ANCESTORS

GT654 and GT480 each could be considered an intermediate common ancestor of the other. This supports the IBD claim.

PHASED MATCHING

None of the genotypes are phased, so no phased matching was possible.

GENERATIONAL MATCHING

There is no lineal descendancy in the triangulated group, so generational matching cannot be tried.

CLOSE RELATIVE MATCHING

GT654 and GT480 are second cousins, and both share this match, supporting the idea that the matching segment was received IBD. For GT611 and GT610, this is a paternal match, and all their close relatives that are genotyped are maternal relatives; genotyped descendants do not match this segment.

MATCH STABILITY

There is no lineal descendancy in the triangulated group, so match stability cannot be evaluated.

ℰ (1 OF 2)

This instance of ℰ is between GT611 (and GT610) and GT654 and GT480.

COMMON ANCESTOR UNIQUENESS

		GT611		GT654	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	6	75%
5	16	16	100%	12	75%
6	32	32	100%	13	41%
7	64	64	100%	14	22%
8	128	108	84%	16	13%
Total	255	235	92%	68	27%

Figure 46: Compiled genealogy completeness evaluation for GT611 and GT654.[321,322]

		GT611		GT480	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	32	100%
7	64	64	100%	55	86%
8	128	108	84%	69	54%
Total	255	235	92%	187	73%

Figure 47: Compiled genealogy completeness evaluation for GT611 and GT480.[323,324]

GT654's genealogy is very incomplete. GT480 is better. But because these two are second cousins, the only portion of the tree that needs to be good is the portion beyond their common ancestor. In GT654's genealogy, the genealogy is complete up to the generation prior to the common ancestor and it is missing half (eight) of the identities in the generation of the common ancestor. GT480's is the same.

GT480's appears to have original research up until the last few generations, then has no sources—a surprise given the thorough documentation prior to that point. GT654's is similar.

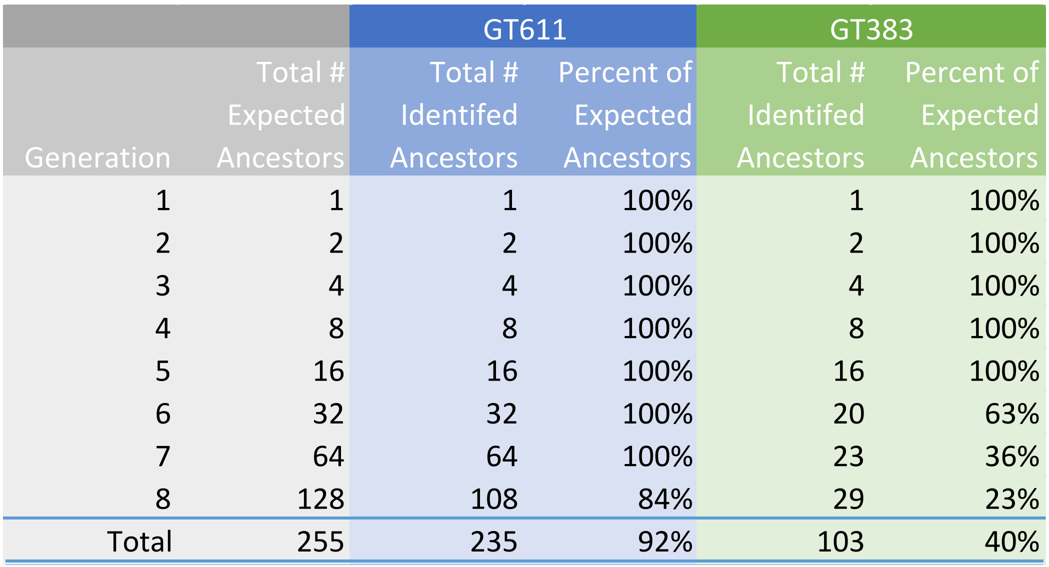
CONCLUSION FOR $\mathcal{E}$ (1 OF 2)

The size of the match with GT480 is a risk because there are no phased comparisons, but the fact that her cousin GT654 has a very likely IBD match with GT611 mitigates this risk. No other analysis factors cast doubt on this match. The match is helped by its fit with the chromosome maps for GT611 and GT610. It is helped by the fact that GT654 and GT480's second cousin relationship acts as an intermediate MRCA. It is hurt by the potential for excessive matching. It is also hurt somewhat by the incompleteness of the genealogies.

$\mathcal{E}$ (2 OF 2)

This instance of $\mathcal{E}$ is between GT611 (and GT610) and GT383.

COMMON ANCESTOR UNIQUENESS



		GT611		GT383	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	20	63%
7	64	64	100%	23	36%
8	128	108	84%	29	23%
Total	255	235	92%	103	40%

Figure 48: Compiled genealogy completeness evaluation for GT611 and GT383.[325,326]

The mechanism that was used to share GT383's genealogy did not include sources; it is essentially an Ahnentafel ancestors report. Again, the genealogy is quite incomplete.

CONCLUSION FOR $\mathcal{E}$ (2 OF 2)

The size of the match with GT383 is a risk because there are no phased comparisons, so there is a small chance that GT383 is not IBD. No other analysis factors cast doubt on this match. The match is helped by its fit with the chromosome maps for GT611 and GT610. It is hurt by the potential for excessive matching. The lineage is also hurt by the incompleteness of the genealogy.

Other Considerations

The MRCA between GT611 and GT383 is one generation newer than the common ancestor shared with GT654 (and GT480). This is not ideal. It could be argued that the MRCA between GT611 and GT383 should be presented as an intermediate MRCA, but this leaves only one instance of $\mathcal{E}$—making a conclusion impossible. But lacking other matches with known genealogies, and lacking any conflict, and given the relative independence of these two instances of $\mathcal{E}$, it seems safe to accept both as instances of $\mathcal{E}$ for the purposes of triangulation until such time as a proper instance of $\mathcal{E}$ presents itself (at which time the MRCA between GT611 and GT383 can become an intermediate MRCA), or until a conflict overturns the proposed solution.

Conclusion Accepting

This shared segment is very "matchy"—an excessive matching risk. So far, only two out of the thirteen possible instances of $\mathcal{E}$ have contributed an answer to the triangulation question—leaving eleven opportunities to introduce conflict. The incompleteness of the genealogies also remains a risk.

The main supporting factors are the strength of GT654's match and the fit with the chromosome maps for GT611 and GT610.

An additional instance of $\mathcal{E}$ that fits the hypothesis would establish a stronger pattern that future instances of $\mathcal{E}$ would confirm the *hypothesis*. Without discarding the *hypothesis*, it seems better not to accept it and pursue additional instances of $\mathcal{E}$.

APPENDIX E: AN X CHROMOSOME MATCH

GT999 shares a large match on the X chromosome with GT793.

KIT #	Chr	START POS	END POS	cM	SNPs
GT793	10	114,905,204	121,216,437	10.6	803
GT793	X	31,676,880	100,037,158	59.7	2,310

Figure 49: Segments that GT999 shares with GT793.[327,328]

The common ancestor shared between GT999 and GT793 was identified as given in Figure 50.

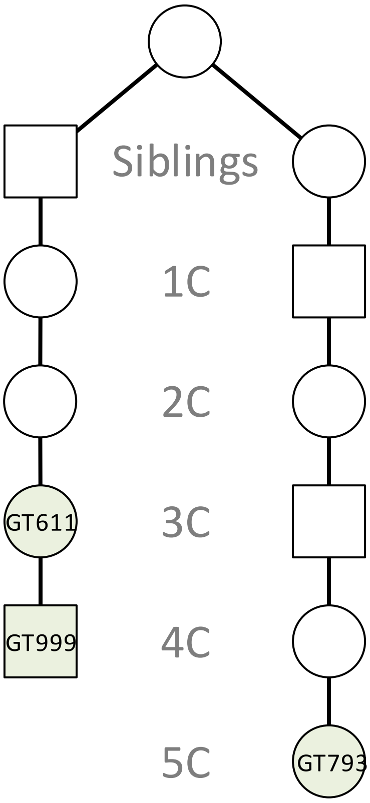


Figure 50: Representation of the lineages that relate GT999 and GT793.[329,330]

Using Figure 3, GT999 is expected to have received (on average) 12.5% of his X chromosome from the specified common ancestor, while GT793 is expected to receive (on average) 6.25% of her X from this ancestor. Using 196.1 cM as the total size of the X chromosome, GT999 is expected to receive about 24.5 cM from this ancestor, and GT793 about 12.3 cM.[331]

So why is the amount of X DNA received by these two cousins from this ancestor—59.7 cM—so high?

The author does not know the answer to this question. Having made this statement, there are anecdotal reports that indicate that the X chromosome may not participate in recombination as often as one might expect, perhaps not reshuffling at all some percentage of the time.[332] In other words, the recombination rate (the number of crossover events expected) is lower for the X chromosome than for the autosomes. Schaffner gives the rate as two thirds of the genome average.[333] The implication of this is that the theoretical percentages given in Figure 3 may be misleading. The percentages are based on the autosome's recombination model, but that model may not fit the realities of the X chromosome.

APPENDIX F: TRIANGULATION FOR GT999 ON CHR1 FROM 180M TO 195M

The following triangulated group was identified for GT999 on Chr1:

KIT #	START POS	END POS	cM	SNPs
GT611	72,017	247,169,190	281.5	54,743
GT709	111,732,674	194,928,236	60.8	13,486
GT610	114,435,751	200,890,245	64.9	13,883
GT861	99,175,785	194,953,541	73.8	16,350
GT901	152,790,212	194,928,236	45.5	9,809
GT553	175,724,793	194,928,236	14.7	3,669
GT439	179,239,541	194,925,386	12.4	2,851
GT436	180,141,290	194,953,541	11.2	2,648

Figure 51: Triangulated group with GT999 for a matching segment on Chr1 from 180M to 195M.[334]

The relationships shared among these genotypes are shown in Figure 52.

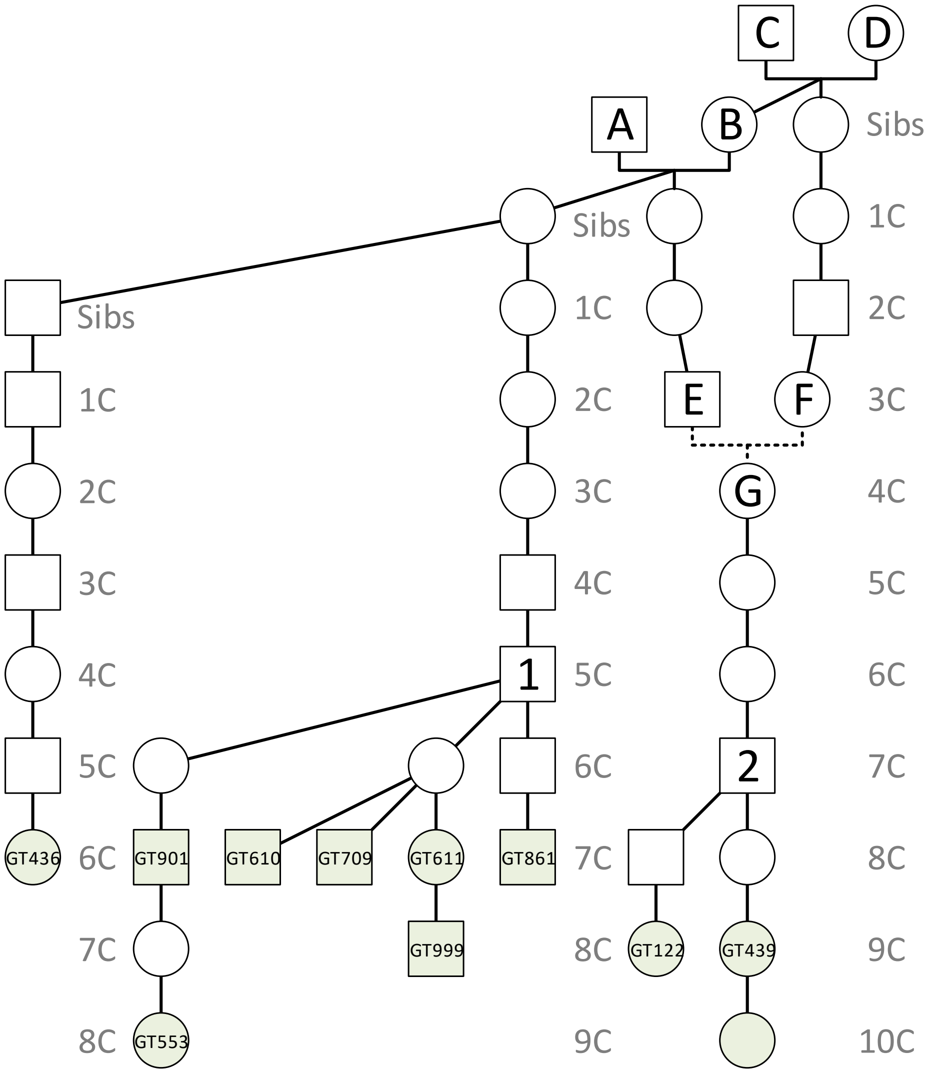


Figure 52: Representation of the known lineages connecting members of a triangulated group for Chr1 (180M to 195M).[335,336,337]

The triangulated group features two family groups and one individual. The first family group (the common ancestor for the family group is marked with a *1*) has representation from seven genotypes (GT611, GT709, GT610, GT861, GT901 and GT553). The second family group (the common ancestor for the family group is marked with a *2*) is represented in the analysis by only one individual: GT439. The GT122 genotype is not part of the triangulated group (because it does not triangulate with GT436).[338] There is another genotype without an identifier—the daughter of GT439—who is reported to match the group, but whose genotype has not been made available for comparisons.[339,340]

There is another genotype from 23andMe™ that has not been represented, but that has been shown to match the triangulated group.[341] A common ancestor with this match has not been identified. Attempts to recruit the cooperation of this match (other than the default genotype sharing on 23andMe™) have been unsuccessful. This genotype is not included in this analysis.

Using the matches listed in Figure 51, the shared match is positioned from 180,141,290 (180M) to 194,925,386 (195M) with a physical length of 14,784,096 bps (14.8 Mbp) and a genetic size of 11.1874 cM.[342]

PROPOSED COMMON ANCESTOR

A search for a common ancestor is usually required to derive the full genealogical benefit of a matching segment. In the case of GT999 and GT439, a lengthy search for a common ancestor ensued.

GT439 had a "brick wall" ancestor identified as Susan SHAW (labeled G in Figure 52) who was married to John J TRAVER and who died in Potsdam, St Lawrence, New York on 18 March 1855.[343] After an extended search of the ancestors of GT999 and GT439, Susan SHAW was identified as an ancestor of interest because GT999 was known to have ancestors that had lived in Potsdam. Continued analysis revealed that Abiel SHURTLEFF (labeled A in Figure 52) and Lydia BARNES (labeled B), the 8th great grandparents of GT999, had a descendant identified as Daniel SHAW (labeled E), who had died in Potsdam and whose age and family might accommodate a daughter that fit what was known about Susan SHAW.[344] To date, all searching for a familial connection between GT999 and GT439, other than the SHAW connection, has been unsuccessful. With atDNA suggesting a family link and given that extensive searching had not revealed any other options, Abiel and Lydia were proposed as common ancestors.

Later, it was shown that GT436 was also related to Abiel and Lydia (as shown in Figure 52).

TRIANGULATION EVALUATION

Before accepting the *hypothesis* that GT999, GT439 and GT436 received the specified matching segment IBD from Abiel and Lydia, the hypothesis must be scrutinized. Factors that affect reliability or that increase the likelihood of errors or misinterpretations must be considered. The items of *information* and *evidence* should agree with each other and with the principles of inheritance. The testing that follows will help determine whether the *hypothesis* can be accepted as a *conclusion*.

TESTS OF ANALYSIS

MATCHING SEGMENT SIZE

Using Figure 19, the probability that any of these matching segments is IBC is less than 5%. As the comparisons were made with a phased genotype, the likelihood that these matches are IBC has been essentially eliminated. All are well above the minimum 5.0 cM threshold suggested for phased matches, the threshold necessary to achieve an acceptable level of false positives.[345] The matching segments are certainly good candidates to be IBD.

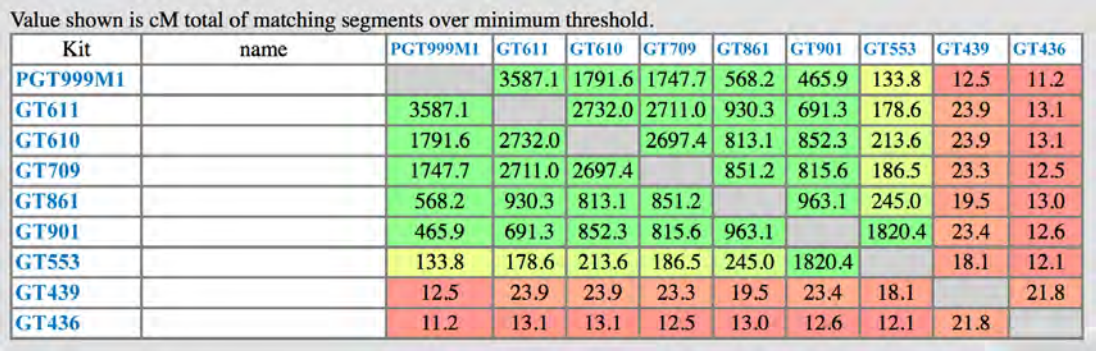
TOTAL SHARED IBD

A comparison with a theoretical average for total sharing is rendered meaningless because the closest relationship (GT436 with GT439) is expected to share just 0.1 cM which is well below the 5 cM segment reliability threshold.[346] The relationship is also outside the set of relationships reported by the *Shared cM Project*.[347,348]

Instead, total IBD sharing testing degrades to an evaluation of the shared segment within the context of Speed and Balding's work.[349] The common ancestor is at G=8 or G=9 (Figure 22) for the member of this triangulated group. A 14.8 Mbp shared region falls in the tranche between 10 and 20 Mbp (Figure 23). There is just under a 40% chance that a segment of this size was shared IBD by an ancestor in the most recent ten generations, and a roughly 5% chance it came from either a G=8 or G=9 ancestor. It would not be considered impossible, or even improbable, that a shared segment of this size (if present) came from the proposed common ancestor IBD.

Tests of Correlation

INDEPENDENCE



Value shown is cM total of matching segments over minimum threshold.

Kit	name	PGT999M1	GT611	GT610	GT709	GT861	GT901	GT553	GT439	GT436
PGT999M1			3587.1	1791.6	1747.7	568.2	465.9	133.8	12.5	11.2
GT611		3587.1		2732.0	2711.0	930.3	691.3	178.6	23.9	13.1
GT610		1791.6	2732.0		2697.4	813.1	852.3	213.6	23.9	13.1
GT709		1747.7	2711.0	2697.4		851.2	815.6	186.5	23.3	12.5
GT861		568.2	930.3	813.1	851.2		963.1	245.0	19.5	13.0
GT901		465.9	691.3	852.3	815.6	963.1		1820.4	23.4	12.6
GT553		133.8	178.6	213.6	186.5	245.0	1820.4		18.1	12.1
GT439		12.5	23.9	23.9	23.3	19.5	23.4	18.1		21.8
GT436		11.2	13.1	13.1	12.5	13.0	12.6	12.1	21.8	

Figure 53: Total amount of atDNA sharing between members of the group that triangulates on a matching Chr1 segment from 180M to 195M.[350]

The report shown in Figure 53 helps evaluate independence within the group. The first seven genotypes in the report are all members of Group 1 (Figure 52). There are two other individuals, independent of Group 1 and independent of each other. Only two independent answers to the triangulation question are possible—the minimum required for triangulation.

EXCESSIVE MATCHING

As discussed previously, there are three independent matches in the triangulated group, and there is only one other candidate that has been identified as matching with the group. This shared segment is not at risk of excessive matching. Rarity strengthens the claim that this matching segment could have been received IBD.

CHROMOSOME MAP CORRELATION

Figure 31 is a Chr1 map for three of the members of the triangulated group: GT611, GT709 and GT610. The matching segment lies between the two yellow lines marking the range 175M and 195M. There are no crossover events within this range. There is only one possible configuration where three siblings could match the same region—the lighter-blue region—and it fits with the siblings that match this triangulated group. The chromosome map agrees with the IBD claim.

INTERMEDIATE COMMON ANCESTORS

There are no known intermediate common ancestors as yet. GT999 has recruited several third, fourth and fifth cousins but none have matched this segment. There are also some AncestryDNA™ matches that might be possible matches based on their shared lineage with GT999, but AncestryDNA™ does not give identifying details (chromosome, start and end locations, size, SNP count) for matching segments, and the genotype administrators have not responded to inquiry.

PHASED MATCHING

GT999 has phased haplotypes that were created with the haplotypes of both parents. All of the genotypes in the triangulated group have been matched against GT999's phased maternal haplotype successfully, essentially eliminating the risk that IBC matches were included in the group—giving a boost to the IBD claim.

GENERATIONAL MATCHING

Using GT999's phased genotype guarantees a generational match with his mother GT611. For GT553, her grandfather GT901 shares the segment. For GT439, her daughter shares the segment. Each of these gives evidence that the matching segment might have been received IBD, giving credibility to the IBD claim.

CLOSE RELATIVE MATCHING

From GT999's point of view, he has five other close relatives that match GT439 and GT436—two uncles, two of his mom's first cousins, and the granddaughter of one of those cousins. There is plenty of evidence that members of Group 1 received this matching segment IBD from their common ancestor (labeled *1* in Figure 52) or his spouse, which gives credence to the idea that they received it IBD from the common ancestor for the triangulated group.

Group 2 includes GT439's first cousin GT122. This cousin does not match GT999 or GT436 but does match the other members of the triangulated group. These matches strengthen the claim of relatedness, and in so doing, strengthen the claim that the matching segment was received IBD.

MATCH STABILITY

There are three pairs of individuals who have a lineal relationship with each other that share this matching segment.

Without the data from her daughter, stability cannot be considered for GT439 and her daughter.

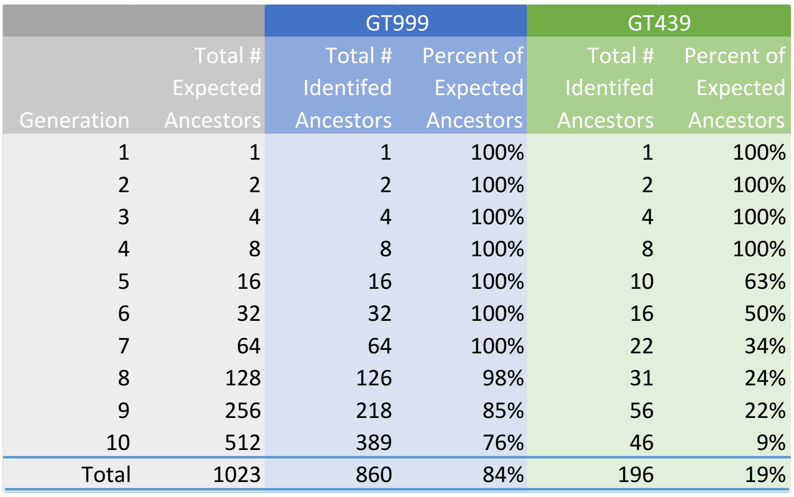
The match is stable—genetically correct; i.e., the newest generation received the matching segment that was the same as or smaller than the older generation's, and the newer generation's matching segment was bounded at or within the boundaries of the older generation's—as it is passed from mother (GT611) to son (GT999).[351,352] This was also true for the matching segment as it was passed from GT901 (grandfather) to GT553 (granddaughter).

A lack of stability (not the case here) would have cast doubt on the IBD claim.

$\mathcal{E}$ (1 OF 2)

The first instance of $\mathcal{E}$ is between Group 1 and Group 2. All of the members of these two family groups match each other in the expected ways (as measured by total sharing).[353,354]

Common Ancestor Uniqueness



		GT999		GT439	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	10	63%
6	32	32	100%	16	50%
7	64	64	100%	22	34%
8	128	126	98%	31	24%
9	256	218	85%	56	22%
10	512	389	76%	46	9%
Total	1023	860	84%	196	19%

Figure 54: Compiled genealogy completeness evaluation for GT999 and GT439.[355,356]

GT999 is used as a proxy for Group 1. Because of close-relative matching, the match is known to have come from GT999's maternal grandmother's father's family—narrowing the portion of his pedigree that needs to be searched to 1/8th of his ancestry. In this portion of his pedigree, only ten ancestors have not been identified—just less than 16% of the expected number of ancestors in this part of the pedigree.

GT439 is used as a proxy for Group 2. Using close-relative matching and ethnicity information, it is believed that this match would be a maternal match for GT439. The pedigree for GT439 is quite incomplete, but the biggest gaps in the pedigree are paternal gaps. The incompleteness is further manifested in the fact that the proposed common ancestor would also “break down a brick wall” for GT439. Because of the gaps in this compiled genealogy, the search for a MRCA had to be supplemented heavily with other published genealogies and supplemental research.[357,358,359] These supplemental activities combined with many hours of searching should serve to mitigate the risks presented by the missing genealogy.

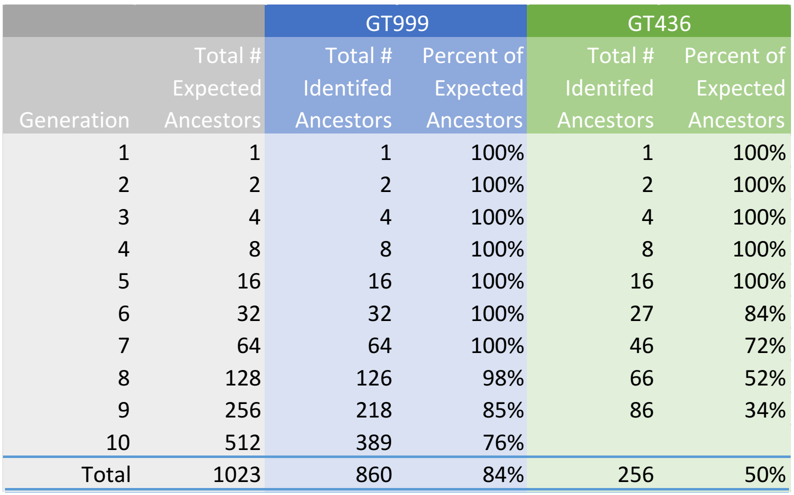
Conclusion for $\mathcal{E}$ (1 of 2)

The main flaw with the first instance of $\mathcal{E}$ is the incompleteness of the genealogies. Much has been done to mitigate this risk by using other published genealogies and doing supplemental research. The author believes that the remaining risk does not leave the solution any more vulnerable than if the pedigrees had been more acceptably complete.

$\mathcal{E}$ (2 OF 2)

The second instance of $\mathcal{E}$ is between Group 1 and GT436.

Common Ancestor Uniqueness



		GT999		GT436	
Generation	Total # Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors	Total # Identifed Ancestors	Percent of Expected Ancestors
1	1	1	100%	1	100%
2	2	2	100%	2	100%
3	4	4	100%	4	100%
4	8	8	100%	8	100%
5	16	16	100%	16	100%
6	32	32	100%	27	84%
7	64	64	100%	46	72%
8	128	126	98%	66	52%
9	256	218	85%	86	34%
10	512	389	76%		
Total	1023	860	84%	256	50%

Figure 55: Compiled genealogy completeness evaluation for GT999 and GT436.[360,361]

The pedigree for GT436 has a number of gaps beyond the fifth generation. The pedigree seems to be a mix of original research and information copied from other compiled genealogies—though in cases of reliance on other's research, it is often supported with well-regarded secondary sources, or one or two original records.

CONCLUSION FOR $\mathcal{E}$ (2 OF 2)

Again, the declaration of a common ancestor (a) is exposed to risk relating to an incomplete genealogy. Not as much has been done to directly mitigate the risks introduced by this second genealogy. Some mitigation is had in the fact that there are two instances of $\mathcal{E}$ (each with independent instances of m and independent lineages to a), which is rare in and of itself; this tends to offset the risk that lingers due to the lack of coverage in the MRCA search.

The match (m) in the second instance of $\mathcal{E}$ bears up to scrutiny; and while there is lingering risk regarding common ancestor (a) due to the incomplete genealogies used in the search for the MRCA, the author feels it is an acceptable risk.

SOLUTION PREDICTS RELATIONSHIPS

Attempts to contact GT436 via GEDmatch in order to find a common ancestor with the group initially failed. Needing an additional match to triangulate with Group 1 and Group 2, the author spent time identifying matches that could potentially corroborate the common ancestor that had been identified between Group 1 and Group 2. He identified an AncestryDNA™ match where the common ancestor identified by AncestryDNA™ was one generation short of the common ancestor, and attempted contact. It turned out that the AncestryDNA™ genotype and the GEDmatch genotype were GT436's genotype. In this sense, the common ancestor identified between Group 1 and Group 2 helped predict that the AncestryDNA™ genotype would be a fit in the triangulated group.

OTHER CONSIDERATIONS

The MRCA between Group 1 and GT436 is one generation closer than the common ancestor shared with Group 2. This is not ideal. It could be argued that the MRCA between Group 1 and GT436 should be represented as an intermediate MRCA, but this leaves only one instance of $\mathcal{E}$—making a conclusion impossible. Lacking other matches with known genealogies, and lacking any conflict, and given the relative independence of these two instances of $\mathcal{E}$, it seems safe to accept both as instances of $\mathcal{E}$ for the purposes of triangulation until such time as a proper instance of $\mathcal{E}$

presents itself (at which time the MRCA between Group 1 and GT436 can become an intermediate MRCA), or until a conflict overturns the proposed solution.

The above discussion has considered the MRCA (labeled *A* and *B* in Figure 52). Interestingly, there is a common lineage between members of this group that could go through a different couple (labeled *C* and *D*) back one generation further. Genetic genealogists assume the MRCA to be the ancestor that contributed a matching segment, but additional matches could change aspects of this assumption. For example, an intermediate common ancestor sharing this matching segment along the lineage between Susan SHAW (labeled *G*) and the couple labeled *C* and *D* would cause us to modify our conclusion about how GT439 received this matching segment. [Whether the matching segment was received by GT439 via Abiel and Lydia (the couple labeled *A* and *B*) or by the couple labeled *C* and *D* would not change the conclusion about the parents of Susan SHAW.]

Conclusion Accepting

None of the testing of the matches (instances of m) cast any doubt in the IBD claim. Some might see lingering doubts in the deficiencies of the search for a MRCA, but the author feels that the work to mitigate these risks, particularly the work leading to the identification of the MRCA between GT999 (Group 1) and GT439 (Group 2), is sufficient to alleviate these doubts. Two lineages were identified by which GT439 could have received this matching segment IBD, but information accumulated thus far is insufficient to select one over the other; for now, the selection of lineage is based on the prevailing MRCA assumption. It is therefore *concluded* that one person in the ancestral couple identified as common to this triangulated group (Abiel SHURTLEFF and Lydia BARNES—labeled *A* and *B* in Figure 52) must have contributed this matching Chr1 segment (180M to 195M) to the members of this triangulated group.

The SHAW Connection

The proposed common ancestor's viability rests (in part) on the following conclusion: that Susan SHAW (married to John J TRAVER) is the daughter of Daniel SHAW and Mary BARROWS.

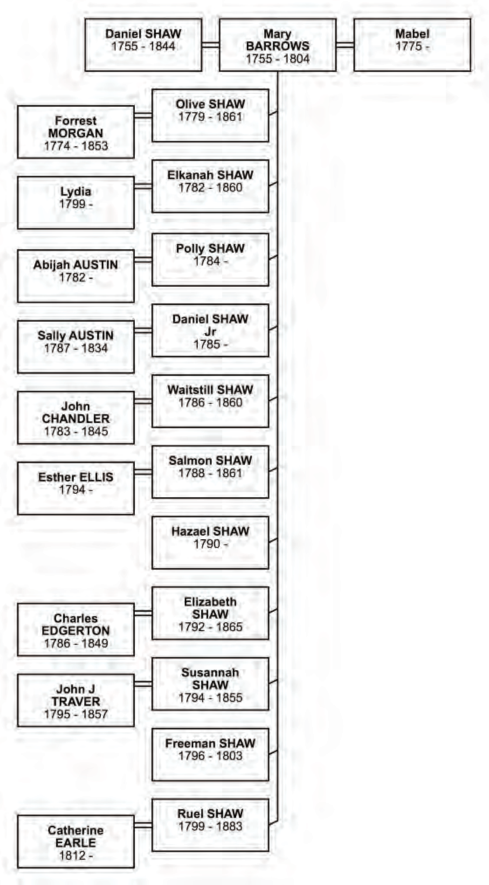


Figure 56: The Daniel SHAW family as discussed in this appendix.

This appendix considers the evidence available about Daniel SHAW and his family and whether Susan is a viable candidate as a daughter in this family. Ultimately, the conclusion outlined here rests on the following:

- A conclusion that the atDNA in common between GT999 and GT439 is shared identical by descent (IBD) through Abiel SHURTLEFF or Lydia BARNES as the proposed common ancestor between GT999 and GT439—a conclusion examined above and incorporated here as the connection to the Daniel SHAW family is considered.
- A conclusion that the persons with the SHAW surname found in the Potsdam arrival record(s)—considered below—are a single family group: Daniel SHAW and his children.
- A conclusion that any persons that resided in Potsdam and bore the SHAW surname in the early 1800s are part of a single family group: Daniel SHAW and his descendants—considered below.

© 2021 This work is licensed under a CC BY-NC-SA 4.0 license.

Daniel SHAW Identified

From a Revolutionary War pension file, personal testimony from him and his wife Mabel reveals the following details about Daniel SHAW (unless otherwise noted).[362] Daniel SHAW was born 12 Oct 1753 in Middleborough, Plymouth, Massachusetts. Sometime during the war, he moved to Plympton, Plymouth, Massachusetts—probably about the time he married Mary BARROWS (a native of Plympton) in Plympton on 6 Aug 1778.[363] Daniel lived in Plympton for eight to ten years, then removed from thence to Bridgewater, Windsor, Vermont. After six years in Bridgewater, he moved to Rochester, Windsor, Vermont. His wife Mary died in Rochester 4 May 1804.[364] Daniel was married a second time in late 1804 to Mrs. Mabel Easton.[365] After twenty years in Rochester, Daniel SHAW moved from Vermont to Potsdam, St Lawrence, New York—maybe in late 1811, as his name was published in the *List of Letters* in five successive issues of the newspaper in early 1812.[366] Daniel died 22 Mar 1844.

The Union

Daniel's sons Elkanah and Salmon were the first of the SHAW family to move from Vermont to Potsdam.[367] A man named William BULLARD convinced several associates (mostly from Royalton, Windsor, Vermont) to join him in an experiment in communal living. Mr. BULLARD was said to be a student in community theory and had published a pamphlet espousing his ideologies. In 1803, Mr. BULLARD identified lands just north of Potsdam Village as a suitable site for the inauguration of his experiment. On 28 Nov 1804, he and his associates took possession of some 2400+ acres of land. In some sources, Elkanah and Salmon are named among Mr. BULLARD's founding associates.

Mr. BULLARD's cooperative organized itself as "The Union", formally adopting a constitution in 1807.[368] The community is said to have thrived well enough. But problems among community participants soon surfaced and, just three years after its formal organization, "The Union" was amicably dissolved. Union lands were divided among the participants. Both Elkanah and Salmon SHAW were among the families receiving lands as "The Union" dissolved, both taking possession of tracts of land on 30 Nov 1810.[369,370]

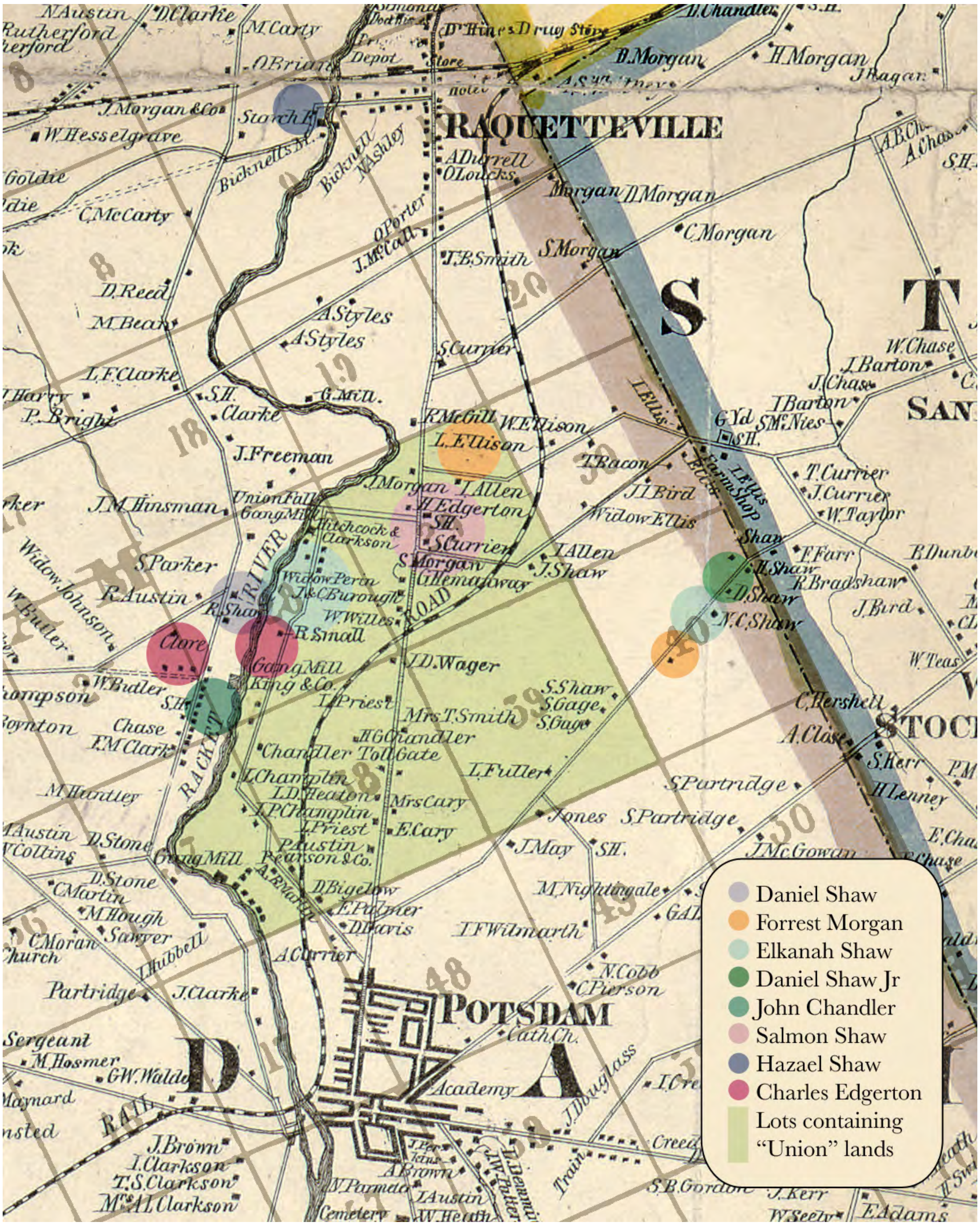


Figure 57: The numbered mile-square lots that contained "The Union" land purchases and the land purchase made by the SHAW family between 1810 and 1818 are overlaid on an 1858 map of St Lawrence county.[371,372,373,374]

Over the next eight years, Daniel, four men believed to be his sons, and three men believed to be his sons-in-law, all purchased lands in the Potsdam area. The above figure shows an approximate location for each of these purchases. The purchases are shown both in the context of the original "Union" land purchase, and in the context of an 1858 map of those same lands. In 1858, the landowners shown are generally contemporaries with Daniel's grandchildren.

In many cases, the descendants of these purchasers are still resident on the purchaser's lands in 1858. The SHAWs' connection to "The Union" is evident. It is easy to see how the proximity of these families to each other suggest they are a family group. These purchases also establish the SHAW presence in Potsdam in the early 1800s—facts important to conclusions drawn in subsequent discussion.

NY 1815 PORT ARRIVALS

There is an index of *NY 1815 Port Arrivals* that includes entries for nine individuals with the surname SHAW (returned in alphabetical order by the search software): Daniel, Daniel Jr., Elizabeth, Elkanah, Hazel, Olive, Salmon, Susanna, and Waitstill.[375] The location given in each index entry is Potsdam, St Lawrence, New York. Given that the stated record type is an arrival record and given that these nine people are the only entries in this entire index with the surname SHAW, and given their arrival in what must have been a small Potsdam port, it seems reasonable to conclude that these individuals are members of a family group, and that perhaps they may have even been recorded in a single arrival record. The index does not identify the original source record(s) used in creating this index, so these assumptions could not be verified. However, it can be shown that Daniel's family is (in part) composed of individuals with these names.

DANIEL SHAW

It is believed that the Daniel in the *NY 1815 Port Arrivals* is Daniel (b. 1753). In support of this belief, it will be shown that the other persons with the SHAW surname in this index fit as children of Daniel. Also, the presence of a Daniel Jr. in the grouping suggests Daniel to be the senior to at least Daniel Jr.

OLIVE SHAW

Olive MORGAN (born about 1779 in Massachusetts) is enumerated with her husband Forrest MORGAN in 1850.[376] She is enumerated in the household of her son Joseph MORGAN in 1860.[377] The marriage record of her son Joseph identifies his parents a Forest MORGAN and Olive SHAW.[378] Forrest MORGAN, Olive's husband, purchased land near the Potsdam SHAW's in 1816 and 1818.[379,380] She was living near her brother Rewell (Ruel) and her son Joseph in 1850.[381] She was living near her brother Salamon (Salmon) in 1860. Her husband, Forrest, was born in Massachusetts, lived in Rochester, Windsor, Vermont, and then lives in Potsdam, St Lawrence, New York—following a similar pattern of migration to that of Daniel himself.[382] Olive is a good fit as a daughter of Daniel SHAW.

ELKANAH SHAW

Elkanah SHAW was born in about 1782.[383] As detailed above, Elkanah and his brother Salmon were founding members of "The Union" with William BULLARD. In 1820, Elkanah was enumerated next to his brother Daniel Jr., and was living near his brother Salmon (with Salmon being enumerated next to William BULLARD).[384] The name Elkanah was very likely given in remembrance of Daniel's father, further evidence that Elkanah is the son of Daniel.[385] Elkanah is a good fit as a son of Daniel SHAW.

DANIEL SHAW JR.

Daniel SHAW Jr. was born in about 1785 in Massachusetts.[386] Daniel Jr. married Sally Austin on 18 Mar 1804 in Rochester, Windsor, Vermont.[387] The marriage was performed by his father Daniel SHAW, a Justice of the Peace—a position his father held from at least 1802 to 1811.[388,389] In 1810, Daniel Jr. was enumerated next to an Abijah Austin (believed to be both his wife's brother and his sister Polly's husband).[390,391] Daniel Jr. purchased land in 1818 near his other siblings, with his land being nearest to his brother Elkanah and brother-in-law Forrest MORGAN (as shown above).[392] In 1820, Daniel Jr. was enumerated next to his brother Elkanah and near his brother Salmon.[393] His designation as Daniel SHAW Jr. would also denote him to be the son of a Daniel SHAW. Daniel Jr. is a good fit as a son of Daniel SHAW.

WAITSTILL SHAW

Waitstill CHANDLER (born in about 1786 in Massachusetts) and her son Nelson were enumerated in 1850 in the household of her married daughter Naomi Pero (Perro).[394] Four years earlier, Waitstill was designated the administratrix of John CHANDLER, suggesting her to be his widow.[395] Waitstill's tombstone connects her to her husband John and reveals that her maiden name was SHAW.[396]

John CHANDLER purchased land near the SHAW family in 1811, with his land being nearest to brother-in-law Charles EDGERTON (Waitstill's sister Elizabeth's husband) and father-in-law Daniel.[397] John was enumerated near several SHAW relations in the 1820, 1830, and 1840 censuses: Charles EDGERTON, Daniel, and Waitstill's brothers Salmon, Daniel Jr., and Elkanah in 1820; Charles EDGERTON, Daniel SHAW, and Forrest MORGAN (Waitstill's sister Olive's husband) in 1830; Charles EDGERTON, Daniel SHAW and Ruel SHAW (Waitstill's brother) in 1840.[398,399,400]

In 1860, Waitstill was again in Naomi's household; a MORGAN (perhaps a relative of her sister Olive's husband?) was a servant in the household.[401]

It is likely that Waitstill received her name in remembrance of her aunt (and father's sister) Waitstill who died 18 Jun 1781 at 11 years of age, approximately 5 years before she was born.[402,403]

The above facts considered together make Waitstill SHAW a solid fit as a daughter of Daniel SHAW.

SALMON SHAW

Salmon SHAW was born in about 1788.[404,405] The sources that give information about his place of birth conflict—one giving Massachusetts, and the other giving Vermont.[406,407] It would seem that his birth is very near the time his family transitioned from Massachusetts to Vermont. An informant without actual knowledge of his birth may have used a knowledge of the family's movements to estimate his birth location and, therefore, introduced this inconsistency.

As detailed above, Salmon SHAW and his brother Elkanah were founding members of "The Union" with William BULLARD. In 1820, Salmon was enumerated next to William BULLARD and near his brothers Elkanah and Daniel Jr.[408] Salmon was enumerated on the same page as his sister Olive's husband Forrest MORGAN in 1830.[409] He was near his nephew Freeman SHAW in 1840.[410] He was living near his sister Olive in 1860.[411]

Salmon named one of his son's Elkanah—a name that connects him to his brother, and his paternal grandfather.[412,413,414]

Considered together, the facts make Salmon SHAW a fit as a son of Daniel SHAW.

Hazael SHAW

Hazael SHAW was born in about 1790 at Bridgewater, Windsor, Vermont—a place Daniel was known to have lived.[415] Hazael was among the SHAW men purchasing land between 1810 and 1818, making a purchase in August of 1817.[416]

Daniel SHAW Jr. named one of his sons Hazael in 1819, an indication from Daniel Jr. that he is related to Hazael, and therefore Hazael's connection to the Daniel SHAW family.

The author wonders if Daniel Jr.'s use of the name Hazael in 1819 is an indication that his brother Hazael is recently deceased. In the search for SHAW records, the last record showing Hazael to be living was his land purchase of 23 Aug 1817.[417] Searches for him in subsequent censuses, newspapers, and in other general searches have not been successful. Also, it can be shown that Daniel Jr. had previously named a son after a recently-deceased brother: his son Freeman after his brother of the same name.[418,419,420]

Given these factors together, Hazael fits as a son of Daniel SHAW.

Elizabeth SHAW

Elizabeth EDGERTON was born in about 1792 in Vermont.[421,422] In both 1850 and 1860, Elizabeth was enumerated in the household of Joseph and Louisa MORGAN.[423] In 1850, the MORGAN household also included Ransom G EDGERTON. In 1860, the MORGAN household included Olive MORGAN (i.e., Olive SHAW, mother of this Joseph MORGAN); the family was also just a few households away from Salmon SHAW.

In the surrogate court's order to grant letters of administration for the estate of Charles EDGERTON, Elizabeth was named Charles's widow, and Ransom his only [living] son.[424] In a similar grant in Ransom's probate proceedings, Louisa MORGAN was named as a sister to Ransom.[425] Elizabeth is the widow of Charles Edgerton and is living in the home of her daughter Louisa.

Elizabeth's presence in the MORGAN household with Olive SHAW is evidence of a family relationship with the SHAW family. If Elizabeth is a daughter of Daniel SHAW (and it appears that she is), she and Olive are sisters, and Olive's son Joseph and Elizabeth's daughter Louisa are first cousins. It is perhaps unusual to find two mothers-in-law residing in the same household with their children; when considered as sisters, it seems a bit more sensible.

Elizabeth's husband Charles purchased land near the SHAWs—probably adjacent to Daniel SHAW.[426,427] Charles was enumerated next to Daniel SHAW in the 1820, 1830, and 1840 censuses.[428,429,430] Charles was also near brothers-in-law John CHANDLER, Salmon SHAW, Elkanah SHAW and Daniel SHAW Jr. in 1820, John CHANDLER and Forrest MORGAN in 1830, and John CHANDLER and Ruel SHAW in 1840. Based on enumeration position in the 1850 census' non-population agriculture schedule, his property is situated next to Joseph MORGAN who is next to Ruel SHAW.[431,432] His proximity to the SHAWs, and especially his proximity to Daniel SHAW, suggest that Elizabeth EDGERTON is a member of the SHAW family.

The above information considered together makes Elizabeth a good fit as a daughter of Daniel SHAW.

SUSANNAH SHAW

Susanna TRAVER was born in about 1794 in Vermont.[433,434,435] She is enumerated in 1850 with her husband John and three children, including a daughter Elisa R TRAVER. This daughter Elisa's death record states that she was born 6 Mar 1825 in Potsdam, St Lawrence, New York and that her parents are John TRAVER and Susan SHAW.[436] Susan (Susannah) died 18 Mar 1855 and was buried in Potsdam.[437]

In the 1850 census, Susannah's place of birth was given as New York.[438] This seems to conflict with the fact that Daniel SHAW did not live in New York at the time of her birth. However, her daughter Elisa and her daughter Susan both gave Susannah's place of birth as Vermont when asked about their mother's place of birth.[439,440] It is not known who the informant was in the 1850 census. It is easy to find examples in census records where the informant did not have actual knowledge of the person's place of birth. On the other hand, if the informant in this record was aware of Susan's place of birth, perhaps one possible explanation for such an answer can be found in the dispute between New York and Vermont over jurisdiction of the lands that now make up Vermont—a dispute that was just being resolved within the one or two years prior to Susan's estimated date of birth.[441] With two daughters reporting her place of birth as Vermont, the New York answer becomes the outlier, giving credence to a conclusion that she was born in Vermont.

Susannah's husband John, while in Potsdam, was not enumerated particularly close to any of the other SHAW siblings or parents in the 1830 and 1840 census.[442,443] The nearest SHAW sibling to Susanna in 1850 (in enumeration order) was Waitstill, but there are sixty-seven dwellings enumerated between their entries.[444] The only proximity to suggest a relationship to other Potsdam SHAW families is Susan's own residence in Potsdam. In fact, after many hours of searching, the only record found to link Susan to Daniel SHAW (outside of her being a SHAW in Potsdam) is the *NY 1815 Port Arrivals* record(s).[445]

There is an Elkanah SHAW (b 1766, d. 1850) married to a Susanna that lived in Bridgewater, Windsor, Vermont.[446] This Elkanah is distinct from the Potsdam Elkanah (b. 1782) by both age (about 16 years different) and geography (one being consistently enumerated in Bridgewater—1800 through 1850, while at the same time the other being consistently enumerated in Potsdam—1820 through 1850).[447,448] One must consider whether the Elkanah and Susanna of Bridgewater might be the Elkanah and Susannah in the arrival records. While it may be possible, it does not seem likely. The more likely candidates would be the candidates actually living their lives in the Potsdam area.

Susannah is certainly a candidate to be a daughter of Daniel SHAW.

Arrival Record Informant

Another interesting feature of the entries in the *NY 1815 Port Arrivals* when considered as a family group, is that, based on the births of their children, at least Olive and Waitstill are already married in 1815. Why, then, would they be listed with their maiden surname in an arrival record? If one considers a family group traveling together and being recorded together in an arrival record(s), it seems possible that a single informant provided the information in that record(s). That everyone in the record(s) is listed as a SHAW suggests that the informant knew everyone in the traveling party to be members of the same family group—that they knew everyone to be a SHAW.

ARRIVAL RECORD CONCLUSION

As given above, the index of *NY 1815 Port Arrivals* includes entries for nine individuals with the surname SHAW. It has been show that each entry can be associated with a candidate person resident in the Potsdam area. These same candidates have been shown to be part of a family group as follows: Daniel SHAW (b. 1755) and his children Olive (b. 1779), Elkanah (b. 1782), Daniel Jr. (b. 1785), Waitstill (b. 1786), Salmon (b. 1788), Hazael (b. 1790), Elizabeth (b. 1792) and Susannah (b. 1794). All of these children fit without conflict into a birth order consistent with Daniel's marriage to Mary BARROWS in 1778, that includes children Polly (b. 1784), Freemam (b. 1796) and Ruel (b. 1799), and all were born prior to Mary's last recorded births (sons Freeman and Ruel). Though unable to access the original arrival record(s), these entries, when combined with the evidence presented above, give evidence of a family relationship. This is important because this grouping includes Susannah who has not appeared in other records with her family.

BEING A SHAW IN POTSDAM

In extensive searching, any person known to have been a SHAW and to have spent time as a resident of Potsdam from the time of the foundation of "The Union" through 1850 (and beyond) have all been shown to be related to Daniel SHAW or one of his descendants, or they have not been eliminated as a descendant. In extensive searching, any person known to have been a SHAW, but that cannot be shown to have spent time as a resident of Potsdam in the specified timeframe, has yet to be shown to be a relation of Daniel SHAW. While the search continues for records about the SHAWs of Potsdam, it seems that being a SHAW in Potsdam between 1804 and 1850 is a strong indicator of a relationship with Daniel SHAW.

BRICK WALL CRUMBLING

This appendix has considered evidence about Daniel SHAW, his children, and whether Susan SHAW (who married John J. Traver) can viably be considered a daughter of Daniel SHAW and Mary BARROWS. From a genealogical proof standard point of view, at least two pieces of evidence are needed to make a conclusion.[449] The author presents the following three evidences in concluding that Susan is, in fact, the daughter of Daniel and Mary:

- evidence of membership in the Daniel SHAW family by virtue of Susannah's association with other members of the Daniel SHAW family in the *NY 1815 Port Arrivals* records,
- evidence of membership in the Daniel SHAW family by virtue of Susannah's status as a SHAW that was resident in Potsdam in the early 1800s,
- and evidence that GT439 received a segment of atDNA that is identical-by-descent (IBD) from Abiel SHURTLEFF or Lydia BARNES, received by GT439 via Susan SHAW.

Some might consider the first two evidences to be sufficient to declare the relationship, and two evidences would be sufficient to declare a conclusion. Adding the atDNA evidence strengthens the argument considerably.

As with any proof argument, this argument stands on the strength and use of all available evidence. As new evidence comes to light, this argument may need to be reconsidered and reformed to include new findings. For now, this argument is stable with both a viable paper trail and confirming atDNA evidence.

[1] Oxford Dictionaries. (2016) 'allele.' *Oxford Dictionaries*. http://www.oxforddictionaries.com/us/definition/english/allele : accessed 21 May 2016.

[2] Wikimedia Foundation, Inc. *Assay*. https://en.wikipedia.org/w/index.php?title=Assay&oldid=724116360 : accessed 13 June 2016. Page Version ID: 724116360.

[3] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 35.

[4] Oxford Dictionaries. (2016) 'diploid.' *Oxford Dictionaries*. http://www.oxforddictionaries.com/us/definition/english/diploid : accessed 21 May 2016.

[5] Oxford Dictionaries. (2016) 'genome.' *Oxford Dictionaries*. http://www.oxforddictionaries.com/us/definition/english/genome : accessed 21 May 2016.

[6] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 17.

[7] 23andMe, Inc. *What is the difference between genotyping and sequencing?* http://customercare.23andme.com/hc/en-us/articles/202904600-What-is-the-difference-between-genotyping-and-sequencing- : accessed 21 May 2016.

[8] Illumina, Inc. *HumanOmniExpress BeadChip Kit*. http://www.illumina.com/products/human_omni_express_beadchip_kits.html : accessed 26 March 2016.

[9] Oxford Dictionaries. (2016) 'haploid.' *Oxford Dictionaries*. http://www.oxforddictionaries.com/us/definition/english/haploid : accessed 21 May 2016.

[10] Merriam-Webster.com. (n.d.) 'locus.' *Merriam-Webster Web*. Merriam-Webster. http://www.merriam-webster.com/dictionary/locus : accessed 22 June 2016.

[11] Oxford Dictionaries. (n.d.) 'mash-up.' *Oxford Dictionaries*. Oxford Dictionaries. http://www.oxforddictionaries.com/us/definition/american_english/mash-up : accessed 15 June 2016.

[12] Merriam-Webster.com. (n.d.) 'mash-up.' *Merriam-Webster Web*. Merriam-Webster. http://www.merriam-webster.com/dictionary/mash%E2%80%93up : accessed 15 June 2016.

[13] Dictionary.com. (n.d.) 'meiosis.' *Dictionary.com Unabridged.* Random House, Inc. http://www.dictionary.com/browse/meiosis : accessed 2 June 2016.

[14] Vocabulary.com. *meiosis*. https://www.vocabulary.com/dictionary/meiosis : accessed 13 June 2016.

[15] Griffith, Sue. *DNA Tips, Tools, & Managing Matches*. http://www.genealogyjunkie.net/dna-tips-tools--managing-matches.html : accessed 2 November 2015.

[16] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 630.

[17] ISOGG Wiki. *Autosomal DNA testing comparison chart*. http://www.isogg.org/wiki/Autosomal_DNA_testing_comparison_chart : accessed 7 November 2015.

[18] AncestryDNA™. *AncestryDNA™ Results for [GT999]*. http://dna.ancestry.com/matches/7A657373-35EB-4667-9BE3-521E6B3078AA : accessed 17 November 2015.

[19] Lewis, Ricki. *Direct-to-Consumer Genetic Testing: A New View*. http://blogs.plos.org/dnascience/2012/11/08/direct-to-consumer-genetic-testing-a-new-view/ : accessed 30 May 2016.

[20] ISOGG Wiki. *Autosomal DNA testing comparison chart*. http://www.isogg.org/wiki/Autosomal_DNA_testing_comparison_chart : accessed 7 November 2015.

[21] Full Genomes Corporation, Inc. *Home*. https://www.fullgenomes.com : accessed 31 May 2016.

[22] Wellcome Trust Sanger Institute. *Timeline: History of genomics*. http://www.yourgenome.org/facts/timeline-history-of-genomics : accessed 31 May 2016.

[23] J Craig Venter Institute. *1902: Boveri & Sutton*. http://www.genomenewsnetwork.org/resources/timeline/1902_Boveri_Sutton.php : accessed 30 May 2016.

[24] J Craig Venter Institute. *1953: Crick & Watson*. http://www.genomenewsnetwork.org/resources/timeline/1953_Crick_Watson.php : accessed 30 May 2016.

[25] J Craig Venter Institute. *1977: Gilbert & Sanger*. http://www.genomenewsnetwork.org/resources/timeline/1977_Gilbert.php : accessed 30 May 2016.

[26] J Craig Venter Institute. *1983: Mullis*. http://www.genomenewsnetwork.org/resources/timeline/1983_Mullis.php : accessed 30 May 2016.

[27] Wellcome Trust Sanger Institute. *Timeline: History of genomics*. http://www.yourgenome.org/facts/timeline-history-of-genomics : accessed 31 May 2016.

[28] Wellcome Trust Sanger Institute. *Timeline: History of genomics*. http://www.yourgenome.org/facts/timeline-history-of-genomics : accessed 31 May 2016.

[29] ISOGG Wiki. *Family Tree DNA*. http://isogg.org/wiki/FamilyTreeDNA : accessed 31 May 2016.

[30] J Craig Venter Institute. *2000: The Human Genome*. http://www.genomenewsnetwork.org/resources/timeline/2000_human.php : accessed 30 May 2016.

[31] Lewis, Ricki. *Direct-to-Consumer Genetic Testing: A New View*. http://blogs.plos.org/dnascience/2012/11/08/direct-to-consumer-genetic-testing-a-new-view/ : accessed 30 May 2016.

[32] ISOGG Wiki. *Family Tree DNA*. http://isogg.org/wiki/FamilyTreeDNA : accessed 31 May 2016.

[33] Ancestry.com. *Ancestry.com Launches New AncestryDNA Service: The Next Generation of DNA Science Poised to Enrich Family History Research: Affordable DNA Test Combines Depth of Ancestry.com Family History Database With an Extensive Collection of DNA Samples to Open New Doors to Family Discovery*. http://ir.ancestry.com/releasedetail.cfm?ReleaseID=669964 : accessed 12 April 2016.

[34] Houghton Mifflin Harcourt Publishing Company. (2011) 'genetics.' *American Heritage Dictionary of the English Language*. Houghton Mifflin Harcourt Publishing Company. http://www.thefreedictionary.com/genetics : accessed 2 June 2016.

[35] Jackson Laboratory. *The Difference Between Genetics and Genomics*. https://www.jax.org/genetics-and-healthcare/genetics-and-genomics/the-difference-between-genetics-and-genomics : accessed 2 June 2016.

[36] *Ibid*.

[37] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 36.

[38] *Ibid*. p. 37.

[39] Bettinger, Blaine. *More X-Chromosome Charts*. http://thegeneticgenealogist.com/2009/01/12/more-x-chromosome-charts/ : accessed 20 December 2015.

[40] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 37.

[41] Campbell, Christopher L., Furlotte, Nicholas A., Eriksson, Nick, et al. (2015) Escape from crossover interference increases with maternal age. *Nature Communications*. 6 p. 6260.

[42] Coop, Graham, Wen, Xiaoquan, Ober, Carole, et al. (2008) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science*. 319 (5868). pp. 1395–1398.

[43] Evans, James P. (2008) Recreational genomics; what's in it for you? *Genetics in Medicine*. 10 (10). pp. 709–710.

[44] 23andMe, Inc. *23 And Me TV Commercial*. https://youtu.be/LrtPoke4X2g : accessed 30 January 2016.

[45] AncestryDNA™. *Ancestry.com DNA Kit TV Commercial, 'Family History'*. http://ispot.tv/a/7g1l : accessed 30 January 2016.

[46] Dudley, Joel T & Karczewski, Konrad J. (2013) *Exploring Personal Genomics*. Oxford: Oxford University Press. p. ix.

[47] Genome Research Limited. *Personal genomics: the future of healthcare?* http://www.yourgenome.org/stories/personal-genomics-the-future-of-healthcare : accessed 30 January 2016.

[48] Evans, James P. (2008) Recreational genomics; what's in it for you? *Genetics in Medicine*. 10 (10). pp. 709–710.

[49] Dudley, *op. cit*.

[50] PGP Global Network. *Personal Genome Project: Harvard*. http://www.personalgenomes.org/harvard : accessed 30 January 2016.

[51] Columbia University and New York Genome Center. *DNA Land*. https://dna.land/ : accessed 3 February 2016.

[52] Evans, James P. (2008) Recreational genomics; what's in it for you? *Genetics in Medicine*. 10 (10). pp. 709–710.

[53] Hill, Kashmir. *Cops are asking Ancestry.com and 23andMe for their customers' DNA*. http://fusion.net/story/215204/law-enforcement-agencies-are-asking-ancestry-com-and-23andme-for-their-customers-dna/ : accessed 19 October 2015.

[54] Petrone, Justin. *Ancestry.com Shutters SMGF Database Amid Murder Case Controversy*. https://www.genomeweb.com/applied-markets/ancestrycom-shutters-smgf-database-amid-murder-case-controversy : accessed 19 October 2015.

[55] ISOGG Wiki. *International Society of Genetic Genealogy Wiki*. http://www.isogg.org : accessed 7 November 2015.

[56] ISOGG Wiki. *Special: Request account*. http://isogg.org/wiki/Special:RequestAccount : accessed 5 March 2016.

[57] Lucy Holman Rector. (2008) Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*. 36 (1). pp. 7–22. http://www.emeraldinsight.com/doi/abs/10.1108/00907320810851998 : accessed 5 Mar 2016.

[58] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC.

[59] Dudley, Joel T & Karczewski, Konrad J. (2013) *Exploring Personal Genomics*. Oxford: Oxford University Press.

[60] Foster, Eugene A., Jobling, M. A., Taylor, P. G., et al. (1998) Jefferson fathered slave's last child. *Nature*. 396 (6706). pp. 27–28. http://www.nature.com/nature/journal/v396/n6706/full/396027a0.html : accessed 5 March 2016.

[61] Coble, Michael D., Loreille, Odile M., Wadhams, Mark J., et al. (2009) Mystery Solved: The Identification of the Two Missing Romanov Children Using DNA Analysis. *PLOS ONE*. 4 (3). p. e4838. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004838 : accessed 5 March 2016.

[62] Gannon, Megan. *It's Really Richard: DNA Confirms King's Remains*. http://www.livescience.com/48963-king-richard-iii-dna-confirmed.html : accessed 5 March 2016.

[63] King, Turi E., Fortes, Gloria Gonzalez, Balaresque, Patricia, et al. (2014) Identification of the remains of King Richard III. *Nature Communications*. 5 p. 5631. http://www.nature.com/ncomms/2014/141202/ncomms6631/full/ncomms6631.html : accessed 5 March 2016.

[64] Hill, Richard. (2012) *Finding Family: My Search for Roots and the Secrets in My DNA*. Scotts Valley, California: On-Demand Publishing, LLC.

[65] Lynn Grub, ed. (2015) *The Adoptee Survival Guide: Adoptees Share Their Wisdom and Tools*. Scotts Valley, California: CreateSpace.

[66] ISOGG Wiki. *Success stories*. http://isogg.org/wiki/Success_stories : accessed 5 March 2016.

[67] Board for Certification of Genealogists. *About BCG*. http://www.bcgcertification.org/aboutbcg/index.html : accessed 28 January 2016.

[68] Board for Certification of Genealogists. (2000) *The BCG Genealogical Standards Manual*. Nashville, Tennessee: Ancestry Publishing.

[69] Rose, Christine. (2009) *Genealogical Proof Standard: Building a Solid Case*. 3rd Ed. Revised. San Jose, California: C R Publications.

[70] Merriman, Brenda Dougall. (2010) *Genealogical Standards of Evidence: A Guide for Family Historians*. Toronto: Dundurn Press.

[71] Mills, Elizabeth Shown. (2009) *Evidence Explained: Citing History Sources from Artifacts to Cyberspace*. 2nd Ed. Baltimore, Maryland: Genealogical Publishing Company.

[72] Jones, Thomas W. (2013) *Mastering Genealogical Proof*. [Kindle version]. Arlington, Virginia: National Genealogical Society.

[73] Board for Certification of Genealogists. (2014) *Genealogy Standards*. 50th Anniversary Ed. Nashville, Tennessee: Ancestry.com.

[74] ISOGG Wiki. *Genetic genealogy blogs*. http://isogg.org/wiki/Genetic_genealogy_blogs : accessed 30 January 2016.

[75] International Society of Genetic Genealogy. *International Society of Genetic Genealogy (ISOGG) [Facebook Group]*. https://www.facebook.com/groups/isogg/ : accessed 30 January 2016.

[76] FamilySearch Wiki. *Upcoming Conferences*. https://familysearch.org/learn/wiki/en/Upcoming_Conferences : accessed 30 January 2016.

[77] International Society of Genetic Genealogy. *DNA-NEWBIE: About Group*. https://groups.yahoo.com/neo/groups/DNA-NEWBIE/info : accessed 30 January 2016.

[78] Johnston, Kathy. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 4 January 2016 5:06 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[79] Kennett, Debbie Cruwys. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 8 January, 7:07 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[80] Bartlett, Jim. *CA and MRCA*. http://segmentology.org/2016/01/02/ca-and-mrca/ : accessed 9 January 2016.

[81] Lee, Jason. *Chasing Stovalls: Is 6th Cousin Triangulation Possible?* http://dnagenealogy.tumblr.com/post/137084622683/chasing-stovalls-is-6th-cousin-triangulation : accessed 3 March 2016.

[82] Bettinger, Blaine. (2016) *[Untitled]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 4 January, 5:06 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[83] Johnston, Kathy. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 4 January, 7:05 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[84] Chen, Daphne. *Should scientists be allowed to change DNA to prevent genetic disease?* https://www.ksl.com/?sid=38376617 : accessed 4 February 2106.

[85] Dallas, Kelsey. (2016) Finding God in one of science's biggest debates — genetic editing. *Deseret News*. 13 January. http://national.deseretnews.com/article/16536/finding-god-in-one-of-sciences-biggest-debates-genetic-editing.html : accessed 14 January 2016.

[86] Jolie, Angelina. (2013) My Medical Choice by Angelina Jolie. *The New York Times*. 14 May. http://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html : accessed 11 March 2016.

[87] Marwa. *Dystopia in Gattaca and Discrimination against Genes*. http://gandt.blogs.brynmawr.edu/web-papers/web-papers-3/dystopia-in-gattaca-and-discrimination-against-genes/ : accessed 11 March 2016.

[88] Dudley, Joel T & Karczewski, Konrad J. (2013) *Exploring Personal Genomics*. Oxford: Oxford University Press. pp. 24-33.

[89] Genetic Genealogy Standards Committee. *Genetic Genealogy Standards*. http://www.geneticgenealogystandards.com : accessed 15 February 2106.

[90] Cowan, Crista. (2016) 'DNA Test Results: Handling the Unexpected.' *RootsTech 2016*,. Salt Lake City 6 February 2016.

[91] Janzen, Tim. (2016) 'Using DNA to Solve Genealogical Problems.' *RootsTech 2016*,. Salt Lake City 6 February 2016. http://tinyurl.com/zsukg8d : accessed 6 February 2016.

[92] ISOGG Wiki. *Autosomal DNA statistics*. http://www.isogg.org/wiki/Autosomal_DNA_statistics : accessed 10 December 2015.

[93] Cooper, Kitty Munson. (2015) 'RT3214 - How to do a DNA Triangulation: Case Studies.' *RootsTech 2016*, Salt Lake City 6 February 2016. http://www.rootstech.org/about/syllabus?lang=eng : accessed 1 February 2016.

[94] Bettinger, Blaine. *Visualizing Distributions for the Shared cM Project*. http://thegeneticgenealogist.com/2015/12/23/visualizing-distributions-for-the-shared-cm-project/ : accessed 16 February 2016.

[95] Bettinger, Blaine. *Visualizing Data From the Shared cM Project*. http://thegeneticgenealogist.com/2015/05/29/visualizing-data-from-the-shared-cm-project/ : accessed 16 February 2016.

[96] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44. http://dougspeed.com/wdytya/ : accessed 15 April 2016.

[97] Henn, Brenna M., Hon, Lawrence, Macpherson, J. Michael, et al. (2012) Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PLoS ONE*. 7 (4). p. e34267. http://dx.doi.org/10.1371/journal.pone.0034267 : accessed 15 February 2016.

[98] Thompson, Elizabeth A. (2013) Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*. 194 (2). pp. 301–326. http://www.genetics.org/content/194/2/301 : accessed 13 February 2016.

[99] Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

[100] Donnelly, Kevin P. (1983) The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*. 23 (1). pp. 34–63.

[101] Bettinger, Blaine. *The Shared cM Project*. https://thegeneticgenealogist.com/2015/05/29/the-shared-cm-project/ : accessed 23 June 2016.

[102] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44. http://dougspeed.com/wdytya/ : accessed 15 April 2016.

[103] Jostins, Luke. *How Many Ancestors Share Our DNA?* http://www.genetic-inference.co.uk/blog/2009/11/how-many-ancestors-share-our-dna/ : accessed 4 February 2016.

[104] Coop, Graham. *How much of your genome do you inherit from a particular grandparent?* https://gcbias.org/2013/10/20/how-much-of-your-genome-do-you-inherit-from-a-particular-grandparent/ : accessed 4 February 2016.

[105] Cooper, Kitty Munson. (2015) 'RT3214 - How to do a DNA Triangulation: Case Studies.' *RootsTech 2016*, Salt Lake City 6 February 2016. http://www.rootstech.org/about/syllabus?lang=eng : accessed 1 February 2016.

[106] Bryant, Rebecca. *A Practical Guide to Using Autosomal DNA (atDNA) for Genealogical Purposes*. https://sites.google.com/site/bryantsofrockislandcreek/2-european-heritage/dna-results/autosomal-dna-atdna : accessed 4 November 2015.

[107] Harman-Hoog, Diane. *A Methodology to Identify Relatives with autosomal DNA Test Data*. http://dnaadoption.com/uploads/DNAadoption/DNAadoption_files/General/Methodology_for_Researching_Autosomal_DNA_Results_V3_1-9-2015.pdf : accessed 10 December 2015.

[108] Bettinger, Blaine. *A Triangulation Intervention*. https://thegeneticgenealogist.com/2016/06/19/a-triangulation-intervention/ : accessed 21 Jun 2016.

[109] *Ibid*.

[110] Bartlett, Jim. *How To Triangulate*. http://segmentology.org/2015/05/11/how-to-triangulate/ : accessed 20 February 2016.

[111] *Ibid*.

[112] Foard, Mesa. (2014) *A Methodology: Identifying your Relatives through your atDNA Results*. http://moodle.dnagedcom.com/MoodleClass/General/Methodology_revised.pdf : accessed 31 October 2015.

[113] Harman-Hoog, Diane. *A Methodology to Identify Relatives with autosomal DNA Test Data*. http://dnaadoption.com/uploads/DNAadoption/DNAadoption_files/General/Methodology_for_Researching_Autosomal_DNA_Results_V3_1-9-2015.pdf : accessed 10 December 2015.

[114] Bettinger, Blaine. (2016) *[Untitled]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 4 January, 5:06 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[115] Bartlett, Jim. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 4 January, 8:45 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[116] Johnston, Kathy. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 4 January, 7:05 PM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[117] Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

[118] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 35.

[119] Bettinger, Blaine. *Small Matching Segments – Friend or Foe?* http://thegeneticgenealogist.com/2014/12/02/small-matching-segments-friend-foe/ : accessed 24 November 2015.

[120] Moore, CeCe. *The Folly of Using Small Segments as Proof in Genealogical Research*. http://www.yourgeneticgenealogist.com/2014/12/the-folly-of-using-small-segments-as.html : accessed 28 November 2015.

[121] Janzen, Tim. (2014) 'Discovering and Verifying your Ancestry using Family Finder.' *2014 International Conference on Genetic Genealogy,*. Houston 11 October 2016. http://tinyurl.com/p22ejo4 : accessed 18 February 2016.

[122] Prairielad. (2015) *Pseudo/False Segments under 5cM*. Family Tree DNA Forums > Universal Lineage Testing (Autosomal DNA) > Family Finder Advanced Topics, 3 October 2015 1:50 AM. http://forums.familytreedna.com/showthread.php?t=38586 : accessed 18 February 2016.

[123] Janzen, Tim. (2014) 'Discovering and Verifying your Ancestry using Family Finder.' *2014 International Conference on Genetic Genealogy*,. Houston 11 October 2016. http://tinyurl.com/p22ejo4 : accessed 18 February 2016.

[124] Bettinger, Blaine. *Small Matching Segments – Friend or Foe?* http://thegeneticgenealogist.com/2014/12/02/small-matching-segments-friend-foe/ : accessed 24 November 2015.

[125] *Ibid*.

[126] Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

[127] *Ibid*.

[128] Bettinger, Blaine. *Small Matching Segments – Friend or Foe?* http://thegeneticgenealogist.com/2014/12/02/small-matching-segments-friend-foe/ : accessed 24 November 2015.

[129] Moore, CeCe. *The Folly of Using Small Segments as Proof in Genealogical Research*. http://www.yourgeneticgenealogist.com/2014/12/the-folly-of-using-small-segments-as.html : accessed 28 November 2015.

[130] Illumina, Inc. *HumanOmniExpress BeadChip Kit*. http://www.illumina.com/products/human_omni_express_beadchip_kits.html : accessed 26 March 2016.

[131] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 119.

[132] Browning, Sharon R. & Browning, Brian L. (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*. 12 (10). pp. 703–714.

[133] Janzen, Tim. (2014) 'Discovering and Verifying your Ancestry using Family Finder.' *2014 International Conference on Genetic Genealogy*, Houston 11 October 2016. http://tinyurl.com/p22ejo4 : accessed 18 February 2016.

[134] Janzen, Tim & Aulicino, Emily. (2013) 'Basics of Chromosome Mapping.' *GFO DNA Interest Group Meeting*, Portland, Oregon 27 July 2013. https://dl.dropboxusercontent.com/u/21841126/Basics%20of%20Chromosome%20Mapping.docx : accessed 23 October 2015.

[135] Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

[136] Johnston, Kathy. (2015) *[Response to Prairielad 'Pseudo/False Segments under 5cM' post]*. Family Tree DNA Forums > Universal Lineage Testing (Autosomal DNA) > Family Finder Advanced Topics, 3 October 2015 12:21 PM. http://forums.familytreedna.com/showthread.php?t=38586 : accessed 18 February 2016.

[137] Janzen, Tim. (2014) 'Discovering and Verifying your Ancestry using Family Finder.' *2014 International Conference on Genetic Genealogy*, Houston 11 October 2016. http://tinyurl.com/p22ejo4 : accessed 18 February 2016.

[138] Janzen, Tim & Aulicino, Emily. (2013) 'Basics of Chromosome Mapping.' *GFO DNA Interest Group Meeting*, Portland, Oregon 27 July 2013. https://dl.dropboxusercontent.com/u/21841126/Basics%20of%20Chromosome%20Mapping.docx : accessed 23 October 2015.

[139] Janzen, Tim. (2015) 'Autosomal DNA Chromosome Mapping Workshop.' *Southern California Genealogy Jamboree 2015*,. Burbank 5 June 2015. http://tinyurl.com/pjg4akw : accessed 20 December 2015.

[140] Janzen, Tim. (2015) *RE: Autosomal DNA mapping*. E-mail to Thad Thomas. 8:30 AM. 28 November. thad.thomas.2013@uni.strath.ac.uk.

[141] ISOGG Wiki. *Triangulation*. http://isogg.org/wiki/Triangulation : accessed 19 February 2106.

[142] ISOGG Wiki. *Chromosome mapping*. http://www.isogg.org/wiki/Chromosome_mapping : accessed 22 October 2015.

[143] ISOGG Wiki. *Identical by state*. http://www.isogg.org/wiki/Identical_by_state : accessed 22 October 2015.

[144] Bartlett, Jim. *Does Triangulation Work?* http://segmentology.org/2015/10/19/does-triangulation-work/ : accessed 17 March 2016.

[145] ISOGG Wiki. *Identical by state*. http://www.isogg.org/wiki/Identical_by_state : accessed 22 October 2015.

[146] Janzen, Tim. (c2012-2014) *Chromosome Mapping for Genetic Genealogy*. http://tinyurl.com/canzmsa : accessed 27 November 2015.

[147] Janzen, Tim & Aulicino, Emily. (2013) 'Basics of Chromosome Mapping.' *GFO DNA Interest Group Meeting*, Portland, Oregon 27 July 2013.

https://dl.dropboxusercontent.com/u/21841126/Basics%20of%20Chromosome%20Mapping.docx : accessed 23 October 2015.

[148] Cooper, Kitty. *Kitty Cooper's Blog*. http://blog.kittycooper.com/ : accessed 10 December 2015.

[149] Griffith, Sue. *DNA Spreadsheets!*. http://www.genealogyjunkie.net/dna-spreadsheets.html : accessed 2 November 2015.

[150] AncestryDNA. (2014) *Your AncestryDNA results are in!*. E-mail to Thad Thomas. 4:13 AM. 6 April. info@blueskiff.com.

[151] AncestryDNA. (2015) *Your AncestryDNA results are in!*. E-mail to Thad Thomas. 11:12 AM. 8 April. info@blueskiff.com.

[152] National Genealogical Society. *2010 NGS Family History Conference Program*. https://web.archive.org/web/20100425064229/http://members.ngsgenealogy.org/Conferences/2010Program.cfm : accessed 30 July 2016.

[153] Utah Genealogical Association. *[SLIG] Tracks*. https://web.archive.org/web/20130120204902/http://www.infouga.org/aem.php?lv=p&epg=27 : accessed 30 July 2016.

[154] FamilySearch International. *GEDCOM X and the Genealogical Research Process*. http://www.gedcomx.org/GEDCOM-X-and-the-Genealogical-Research-Process.html : accessed 21 April 2016.

[155] Thomas, Thad. (2015) *Re: Thad Thomas - MSc topic*. E-mail to Alasdair Macdonald and Graham Holton. 2:23 AM. 29 August. thad.thomas.2013@uni.strath.ac.uk.

[156] Jones, Thomas W. (2013) *Mastering Genealogical Proof*. [Kindle version]. Arlington, Virginia: National Genealogical Society. p. 7.

[157] *Ibid*. pp. 23-29.

[158] *Ibid*. pp. 13-16.

[159] *Ibid*. pp. 23-24.

[160] *Ibid. pp.53- 64.*

[161] *Ibid*. p. 64.

[162] *Ibid*. p. 83.

[163] FamilySearch International. *GEDCOM X and the Genealogical Research Process*. http://www.gedcomx.org/GEDCOM-X-and-the-Genealogical-Research-Process.html : accessed 21 April 2016.

[164] 23andMe, Inc. *What is the difference between genotyping and sequencing?* http://customercare.23andme.com/hc/en-us/articles/202904600-What-is-the-difference-between-genotyping-and-sequencing- : accessed 21 May 2016.

[165] Illumina, Inc. *HumanOmniExpress BeadChip Kit*. http://www.illumina.com/products/human_omni_express_beadchip_kits.html : accessed 26 March 2016.

[166] ISOGG Wiki. *Autosomal DNA testing comparison chart*. http://www.isogg.org/wiki/Autosomal_DNA_testing_comparison_chart : accessed 7 November 2015.

[167] Jobling, Mark A, Hollox, Edward, Hurles, Matthew, et al. (2014) *Human Evolutionary Genetics*. 2nd ed. London: Garland Science, Taylor & Francis Group, LLC. p. 17.

[168] J Craig Venter Institute. *Genome Variations*. http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp4_1.shtml : accessed 30 May 2016.

[169] Bartlett, Jim. *Benefits of Triangulation*. http://segmentology.org/2015/05/09/benefits-of-triangulation/ : accessed 9 January 2016.

[170] Hartley, Joel M. *Beware the False DNA Match – Hartley DNA & Genealogy*. http://www.jmhartley.com/HBlog/?p=540 : accessed 13 May 2016.

[171] Estes, Roberta. *Concepts – Identical by…Descent, State, Population and Chance*. https://dna-explained.com/2016/03/10/concepts-identical-bydescent-state-population-and-chance/ : accessed 6 June 2016.

[172] Waldron, Paddy. *Measuring the length, the rarity and the relevance of shared autosomal DNA*. http://www.pwaldron.info/DNA/significance.html : accessed 6 December 2015.

[173] Stevens, Eric L., Heckenberg, Greg, Roberson, Elisha D. O., et al. (2011) Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. *PLoS Genetics*. 7 (9). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178600/ : accessed 14 February 2016.

[174] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44. http://dougspeed.com/wdytya/ : accessed 29 April 2016.

[175] Henn, Brenna M., Hon, Lawrence, Macpherson, J. Michael, et al. (2012) Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PLoS ONE*. 7 (4). p. e34267. http://dx.doi.org/10.1371/journal.pone.0034267 : accessed 15 February 2016.

[176] Mount, Steve. *Genetic Genealogy and the Single Segment*. http://ongenetics.blogspot.co.uk/2011/02/genetic-genealogy-and-single-segment.html : accessed 11 May 2016.

[177] Coop, Graham. *How much of your genome do you inherit from a particular ancestor?* https://gcbias.org/2013/11/04/how-much-of-your-genome-do-you-inherit-from-a-particular-ancestor/ : accessed 4 February 2016.

[178] Gene By Gene, Ltd. *Autosomal Tests*. https://www.familytreedna.com/learn/dna-basics/autosomal/ : accessed 4 June 2016.

[179] Donnelly, Kevin P. (1983) The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*. 23 (1). pp. 34–63.

[180] Jostins, Luke. *How Many Ancestors Share Our DNA?* http://www.genetic-inference.co.uk/blog/2009/11/how-many-ancestors-share-our-dna/ : accessed 4 February 2016.

[181] Bartlett, Jim. *The Porcupine Chart*. http://segmentology.org/2015/08/07/the-porcupine-chart/ : accessed 12 March 2016.

[182] Coop, Graham. *How many genetic ancestors do I have?* https://gcbias.org/2013/11/11/how-does-your-number-of-genetic-ancestors-grow-back-over-time/ : accessed 4 June 2016.

[183] Estes, Roberta. *Ethnicity Testing – A Conundrum*. https://dna-explained.com/2016/02/10/ethnicity-testing-a-conundrum/ : accessed 5 June 2016.

[184] Canada, R A. *How does Population Finder determine the percentages of different ancestries?* https://web.archive.org/web/20140606004944/https://www.familytreedna.com/learn/autosomal-ancestry/ethnic-origins/population-finder-find-ancestry/ : accessed 4 June 2016.

[185] Estes, Roberta. *Ethnicity Testing – A Conundrum*. https://dna-explained.com/2016/02/10/ethnicity-testing-a-conundrum/ : accessed 5 June 2016.

[186] *Ibid*.

[187] GEDmatch.com. *Admixture/Oracle Population Search Utility*. https://www.gedmatch.com/ : accessed 4 June 2016.

[188] Bartlett, Jim. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 5 January, 11:25 AM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[189] AncestryDNA™. *AncestryDNA Matching Help and Tips: Should other family members get tested?* http://dna.ancestry.com/ : accessed 4 February 2016.

[190] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44. http://dougspeed.com/wdytya/ : accessed 29 April 2016.

[191] Ball, Catherine A, Barber, Mathew J, Byrnes, Jake K, et al. (2014) *DNA Circles White Paper: Identifying groups of descendants using pedigrees and genetically inferred relationships in a large database*. Lehi, Utah. http://dna.ancestry.com/resource/whitePaper/AncestryDNA-DNA-Circles-White-Paper : accessed 26 May 2016.

[192] Bettinger, Blaine. *Visualizing Distributions for the Shared cM Project*. http://thegeneticgenealogist.com/2015/12/23/visualizing-distributions-for-the-shared-cm-project/ : accessed 16 February 2016.

[193] Henn, Brenna M., Hon, Lawrence, Macpherson, J. Michael, et al. (2012) Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. *PLoS ONE*. 7 (4). p. e34267. http://dx.doi.org/10.1371/journal.pone.0034267 : accessed 15 February 2016.

[194] Genetic Genealogy Standards Committee. *Genetic Genealogy Standards*. http://www.geneticgenealogystandards.com : accessed 15 February 2106.

[195] AncestryDNA™. Member Matches for an8181: M.G. http://dna.ancestry.com/tests/CB243397-055C-40A8-8F89-18EF5060480B/match/DF7EAF1F-AC01-47BB-8154-BADB98671834 : accessed 2 April 2016.

[196] Bettinger, Blaine. *Visualizing Distributions for the Shared cM Project*. http://thegeneticgenealogist.com/2015/12/23/visualizing-distributions-for-the-shared-cm-project/ : accessed 16 February 2016.

[197] *Ibid.*

[198] [an8181's cousin]. *[an8181's cousin] Family Tree*. http://trees.ancestry.com/tree/9335040/family : accessed 23 January 2016.

[199] Wikimedia Foundation, Inc. *Centimorgan*. https://en.wikipedia.org/wiki/Centimorgan : accessed 2 December 2015.

[200] GEDmatch.com. Result of default 'Matching Segment Search' for GT999. https://www.gedmatch.com/ : accessed 23 May 2016.

[201] [GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 25 December 2015.

[202] [GT831]. *[GT831] Family Tree*. http://trees.ancestry.com/tree/15002490/family : accessed 25 December 2015.

[203] GEDmatch.com. Results of default 'one-to-one' comparison between GT999 and GT831. https://www.gedmatch.com/ : accessed 7 May 2016.

[204] [GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 4 January 2016.

[205] [GT124]. *[GT124] Family Tree*. http://trees.ancestry.com/tree/80075627/family : accessed 4 January 2016.

[206] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT999 and GT124. https://www.gedmatch.com/ : accessed 7 May 2016.

[207] Matise Laboratory of Computational Genetics. *Map Interpolator of the Rutgers Map*. http://compgen.rutgers.edu/mapinterpolator : accessed 3 June 2016.

[208] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT439. https://www.gedmatch.com/ : accessed 7 May 2016.

[209] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT557. https://www.gedmatch.com/ : accessed 7 May 2016.

[210] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT439 and GT557. https://www.gedmatch.com/ : accessed 7 May 2016.

[211] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT831 and GT124. https://www.gedmatch.com/ : accessed 7 May 2016.

[212] Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

[213] Levin, A J. *Downside to DNA, Part 2*. https://thegenealogistdotca.wordpress.com/2016/03/24/downside-to-dna-part-2/ : accessed 12 April 2016.

[214] GEDmatch.com. Result of default ‘Matching Segment Search’ for GT999. https://www.gedmatch.com/ : accessed 23 May 2016.

[215] Durand, Eric Y., Eriksson, Nicholas, & McLean, Cory Y. (2014) Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution*. 31 (8). pp. 2212–2222. http://mbe.oxfordjournals.org/content/31/8/2212 : accessed 27 November 2105.

[216] Bettinger, Blaine. *Small Matching Segments – Friend or Foe?* http://thegeneticgenealogist.com/2014/12/02/small-matching-segments-friend-foe/ : accessed 24 November 2015.

[217] ISOGG Wiki. *Autosomal DNA match thresholds*. http://www.isogg.org/wiki/Autosomal_DNA_match_thresholds : accessed 12 May 2016.

[218] Bettinger, Blaine. *Visualizing Data From the Shared cM Project*. http://thegeneticgenealogist.com/2015/05/29/visualizing-data-from-the-shared-cm-project/ : accessed 16 February 2016.

[219] Bettinger, Blaine. *Visualizing Distributions for the Shared cM Project*. http://thegeneticgenealogist.com/2015/12/23/visualizing-distributions-for-the-shared-cm-project/ : accessed 16 February 2016.

[220] Janzen, Tim. (2016) 'Using DNA to Solve Genealogical Problems.' *RootsTech 2016*,. Salt Lake City 6 February 2016. http://tinyurl.com/zsukg8d : accessed 6 February 2016.

[221] ISOGG Wiki. *Autosomal DNA statistics*. http://www.isogg.org/wiki/Autosomal_DNA_statistics : accessed 10 December 2015.

[222] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT880. https://www.gedmatch.com/ : accessed 28 May 2016.

[223] [GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 25 December 2015.

[224] [GT880]. *[GT880]lineagetree*. http://person.ancestry.com/tree/1825910/person/-1753013398 : accessed 28 May 2016.

[225] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44.

[226] *Ibid*.

[227] Ball, Catherine A, Barber, Mathew J, Byrnes, Jake K, et al. (2014) *DNA Circles White Paper: Identifying groups of descendants using pedigrees and genetically inferred relationships in a large database*. Lehi, Utah. http://dna.ancestry.com/resource/whitePaper/AncestryDNA-DNA-Circles-White-Paper : accessed 26 May 2016.

[228] GEDmatch.com. Result of ‘Multiple Kit Analysis: Autosomal DNA comparison matrix’ for GT999, GT163, GT625, GT177, GT381, GT136, GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 13 May 2016.

[229] [GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 25 December 2015.

[230] [GT978]. *[GT978] Family Tree*. http://trees.ancestry.com/tree/51133039/person/13271713048 : accessed 25 April 2016.

[231] [GT381]. (2016) *Re: MRCA: Joseph Call & Mary Sanderson*. E-mail to Thad Thomas. 6:13 AM. 30 April. thad.thomas.2013@uni.strath.ac.uk.

[232] GEDmatch.com. Result of default ‘GEDmatch DNA Segment Search’ for GT999, PGT999P1 and PGT999M1. https://www.gedmatch.com/ : accessed 17 May 2016.

[233] Ball, Catherine A, Barber, Mathew J, Byrnes, Jake, et al. (2016) *AncestryDNA™ Matching White Paper: Discovering genetic matches across a massive, expanding genetic database*. Lehi, Utah: AncestryDNA™. http://dna.ancestry.com/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf : accessed 17 May 2016.

[234] GEDmatch.com. *Phased data generator*. https://www.gedmatch.com/ : accessed 20 May 2016.

[235] ISOGG Wiki. *Phasing*. http://www.isogg.org/wiki/Phasing : accessed 25 October 2015.

[236] Janzen, Tim. (2014) *Re: [AUTOSOMAL-DNA] Need Help on Triangulated Groups or Phasingand Chromosome Mapping*. AUTOSOMAL-DNA, 15 April 2014 21:49:02 -0700. http://archiver.rootsweb.ancestry.com/th/read/AUTOSOMAL-DNA/2014-04/1397623742 : accessed 20 May 2016.

[237] Bartlett, Jim. *CA and MRCA*. http://segmentology.org/2016/01/02/ca-and-mrca/ : accessed 9 January 2016.

[238] GEDmatch.com. Result of default 'one-to-one' comparison between GT611 and each of GT383, GT480 and GT654. https://www.gedmatch.com/ : accessed 1 April 2016.

[239] GEDmatch.com. Result of default 'one-to-one' comparison between GT383 and each of GT480 and GT654. https://www.gedmatch.com/ : accessed 1 April 2016.

[240] GEDmatch.com. Result of default 'one-to-one' comparison between GT480 and GT654. https://www.gedmatch.com/ : accessed 1 April 2016.

[241] [GT383]. *Ancestors of [GT383]*. http://freepages.folklore.rootsweb.ancestry.com/~fromheretopaternity/index.htm : accessed 1 April 2016.

[242] [GT480]. *[GT480 Family Tree]*. http://trees.ancestry.com/tree/70185236/family : accessed 1 April 2016.

[243] [GT654]. *[GT654] Family Tree*. http://trees.ancestry.com/tree/63112612/family : accessed 1 April 2016.

[244] Bartlett, Jim. *Does Triangulation Work?* http://segmentology.org/2015/10/19/does-triangulation-work/ : accessed 17 March 2016.

[245] Lee, Jason. *Triangulation and the Birthday Problem*. http://dnagenealogy.tumblr.com/post/140323916583/triangulation-and-the-birthday-problem : accessed 3 March 2016.

[246] [GT357]. (2016) *Re: DNA Match*. E-mail to Thad Thomas. 9:56 PM. 25 April. thad.thomas.2013@uni.strath.ac.uk.

[247] GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT118, GT163, GT999 and GT186 with GT208. https://www.gedmatch.com/ : accessed 20 May 2016.

[248] GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT118, GT163, GT999 and GT186 with GT293. https://www.gedmatch.com/ : accessed 20 May 2016.

[249] Lee, Jason. *The Use of Crossover Lines to Determine Segment Matches with Grandparents Among Siblings — Visual Phasing*. http://dnagenealogy.tumblr.com/post/137722603308/the-use-of-crossover-lines-to-determine-segment : accessed 22 February 2016.

[250] GEDmatch.com. Result of default 'one-to-one' comparison between GT611 and each of GT709, GT610 and GT926. https://www.gedmatch.com/ : accessed 14 May 2016.

[251] GEDmatch.com. Result of default 'one-to-one' comparison between GT709 and each of GT610 and GT926. https://www.gedmatch.com/ : accessed 14 May 2016.

[252] GEDmatch.com. Result of default 'one-to-one' comparison between GT610 and GT926. https://www.gedmatch.com/ : accessed 14 May 2016.

[253] ISOGG Wiki. *Identical by descent*. http://isogg.org/wiki/Identical_By_Descent : accessed 13 May 2016.

[254] *Ibid*.

[255] GEDmatch.com. Result of default 'Matching Segment Search' for GT999. https://www.gedmatch.com/ : accessed 23 May 2016.

[256] GEDmatch.com. Result of default 'Matching Segment Search' for GT712. https://www.gedmatch.com/ : accessed 23 May 2016.

[257] Ball, Catherine A, Barber, Mathew J, Byrnes, Jake, et al. *AncestryDNA™ Matching White Paper: Discovering genetic matches across a massive, expanding genetic database*. http://dna.ancestry.com/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf : accessed 17 May 2016.

[258] Edwards, Dan. *Chromosome Pile-Ups in Genetic Genealogy: Examples from 23andMe and FTDNA*. http://ourpuzzlingpast.com/geneblog/2015/01/31/chromosome-pile-ups-in-genetic-genealogy-examples-from-23andme-and-ftdna/ : accessed 23 May 2016.

[259] *Ibid*.

[260] Browning, Sharon R. & Browning, Brian L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 46 (1). pp. 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534 : accessed 13 February 2016.

[261] GEDmatch.com. Result of default multi-kit '2-D Chromosome Browser comparison' and 'Autosomal Matrix comparison' between PGT368P1 and each of GT832, GT365, GT431, GT651, GT463, GT418, GT658, GT588, GT193, GT237, GT833, GT779, GT176, GT385, GT589, GT631, GT726, GT971, GT192, GT310,

GT132, GT112, GT391, GT918, GT594, GT259, GT575, GT111, GT505, GT534, GT432 and GT572. https://www.gedmatch.com/ : accessed 14 May 2016.

[262] [GT594]. *[GT594] Family Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 7 April 2016.

[263] [GT385 administrator]. (2016) *Re: Looking for common ancestor with GT385* .... E-mail to Thad Thomas. 06:21 PM. 8 March. thad.thomas.2013@uni.strath.ac.uk.

[264] [GT779]. (2016) *RE: Looking for common ancestor with [GT779]* .... E-mail to Thad Thomas. 08:05 AM. 8 March. thad.thomas.2013@uni.strath.ac.uk.

[265] Bartlett, Jim. (2016) *[Response to untitled Blaine T. Bettinger post]*. International Society of Genetic Genealogy (ISOGG) [Facebook Group], 5 January, 11:25 AM. https://www.facebook.com/groups/isogg/permalink/10153906822642922/ : accessed 9 January 2016.

[266] Estes, Roberta. *Concepts – CentiMorgans, SNPs and Pickin' Crab*. https://dna-explained.com/2016/03/30/concepts-centimorgans-snps-and-pickin-crab/ : accessed 7 June 2016.

[267] Genetic Genealogy Standards Committee. *Genetic Genealogy Standards*. http://www.geneticgenealogystandards.com : accessed 15 February 2106.

[268] [Husband of an8181]. (2016) *Adoption Information*. E-mail to Thad Thomas. 10:55:12 AM. 23 January. thad.thomas@ldschurch.org.

[269] AncestryDNA™. Member Matches for an8181: C.S. http://dna.ancestry.com/tests/CB243397-055C-40A8-8F89-18EF5060480B/match/6B098128-56DE-441F-A005-3D3661BB65A8 : accessed 2 April 2016.

[270] an8181. *[an8181's] Biological Family Tree*. http://trees.ancestry.com/tree/86926265/family : accessed 25 December 2015.

[271] Bartlett, Jim. *Getting Started with Autosomal DNA Part I*. http://segmentology.org/2015/11/22/getting-started-with-autosomal-dna-part-i/ : accessed 18 March 2016.

[272] Griffith, Sue. *DNA Tips, Tools, & Managing Matches*. http://www.genealogyjunkie.net/dna-tips-tools--managing-matches.html : accessed 2 November 2015.

[273] Cooper, Kitty. *Using your DNA test results: the Basics for Genealogists*. http://blog.kittycooper.com/2015/03/using-your-dna-test-results-the-basics-for-genealogists/ : accessed 10 December 2015.

[274] GEDmatch.com. Result of default 'one-to-one' comparison between PGT999M1 and GT611, GT861, GT709, GT610, GT901, GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

[275] [GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 25 December 2015.

[276] [GT124]. *[GT124] Family Tree*. http://trees.ancestry.com/tree/80075627/family : accessed 4 January 2016.

[277] [GT831]. *[GT831] Family Tree*. http://trees.ancestry.com/tree/15002490/family : accessed 25 December 2015.

[278] Matise Laboratory of Computational Genetics. *Map Interpolator of the Rutgers Map*. http://compgen.rutgers.edu/mapinterpolator : accessed 3 June 2016.

[279] Durand, Eric Y., Eriksson, Nicholas, & McLean, Cory Y. (2014) Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution*. 31 (8). pp. 2212–2222. http://mbe.oxfordjournals.org/content/31/8/2212 : accessed 27 November 2105.

[280] ISOGG Wiki. *Autosomal DNA statistics*. http://www.isogg.org/wiki/Autosomal_DNA_statistics : accessed 10 December 2015.

[281] GEDmatch.com. Result of 'one-to-one' comparison (using 300 SNP and 4 cM match thresholds) between PGT999M1 and each of GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

[282] Janzen, Tim. (2016) 'Using DNA to Solve Genealogical Problems.' *RootsTech 2016*,. Salt Lake City 6 February 2016. http://tinyurl.com/zsukg8d : accessed 6 February 2016.

[283] GEDmatch.com. Result of 'Multiple Kit Analysis: Autosomal DNA comparison matrix' for GT999, GT611, GT861, GT709, GT610, GT901, GT124, GT732, GT681, and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

[284] GEDmatch.com. Result of default 'one-to-one' comparison between GT186 and each of GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

[285] GEDmatch.com. Result of default 'one-to-one' comparison between GT611 and each of GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

[286] GEDmatch.com. Result of default 'one-to-one' comparison between PGT999M1 and GT611, GT861, GT709, GT610, GT901, GT124, GT732 and GT831. https://www.gedmatch.com/ : accessed 16 June 2016.

[287] FamilySearch Family Tree. *Ancestry: KW89-JLZ*. https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[288] [GT124]. *[GT124] Family Tree*. http://trees.ancestry.com/tree/80075627/family : accessed 4 January 2016.

[289] GEDmatch.com. Result of 'one-to-one' comparison (using 300 SNP and 4 cM match thresholds) between PGT999P1 and each of GT163, GT625, GT177, GT381, GT136, GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 1 April 2016.

[290] Matise Laboratory of Computational Genetics. *Map Interpolator of the Rutgers Map*. http://compgen.rutgers.edu/mapinterpolator : accessed 3 June 2016.

[291] Durand, Eric Y., Eriksson, Nicholas, & McLean, Cory Y. (2014) Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution*. 31 (8). pp. 2212–2222. http://mbe.oxfordjournals.org/content/31/8/2212 : accessed 27 November 2105.

[292] ISOGG Wiki. *Autosomal DNA statistics*. http://www.isogg.org/wiki/Autosomal_DNA_statistics : accessed 10 December 2015.

[293] Bettinger, Blaine. *Visualizing Data From the Shared cM Project*. http://thegeneticgenealogist.com/2015/05/29/visualizing-data-from-the-shared-cm-project/ : accessed 16 February 2016.

[294] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44.

[295] GEDmatch.com. Result of 'Multiple Kit Analysis: Autosomal DNA comparison matrix' for GT999, GT163, GT625, GT177, GT381, GT136, GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 13 May 2016.

[296] GEDmatch.com. Result of default 'one-to-one' comparison between GT186 and each of GT978, GT789, GT606 and GT491. https://www.gedmatch.com/ : accessed 9 June 2016.

[297] GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT368, GT859 and GT929 with GT978. https://www.gedmatch.com/ : accessed 20 May 2016.

[298] FamilySearch Family Tree. *Ancestry: KW89-JLZ*. https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[299] [GT978]. *[GT978] Family Tree*. http://trees.ancestry.com/tree/51133039/person/13271713048 : accessed 25 April 2016.

[300] FamilySearch Family Tree. *Ancestry: [Parents of] KWC2-Q3X*. https://familysearch.org/tree/#view=tree&person=KWC2-Q3X&section=fan : accessed 9 June 2016.

[301] [GT381]. (2016) *Re: MRCA: Joseph Call & Mary Sanderson*. E-mail to Thad Thomas. 6:13 AM. 30 April. thad.thomas.2013@uni.strath.ac.uk.

[302] GEDmatch.com. Result of default 'one-to-many' comparison for GT381. https://www.gedmatch.com/ : accessed 9 June 2016.

[303] GEDmatch.com. Result of default 'one-to-one' comparison between GT186 and each of GT381, GT625 and GT177. https://www.gedmatch.com/ : accessed 9 June 2016.

[304] GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT368, GT859 and GT929 with GT177. https://www.gedmatch.com/ : accessed 20 May 2016.

[305] FamilySearch Family Tree. *Ancestry: KW89-JLZ*. https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[306] [GT177]. *Family Research tree (DNA matches)*. http://person.ancestry.com/tree/81972480/person/74006448666 : accessed 10 June 2016.

[307] [GT357]. (2016) *Re: DNA Match*. E-mail to Thad Thomas. 4:08 AM. 3 May. thad.thomas.2013@uni.strath.ac.uk.

[308] GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT241, GT787, GT502 and GT357 with each of GT163, GT978 and GT625. https://www.gedmatch.com/ : accessed 20 May 2016.

[309] GEDmatch.com. Result of default 'one-to-one' comparison between GT186 and GT136. https://www.gedmatch.com/ : accessed 9 June 2016.

[310] GEDmatch.com. Result of default 'one-to-one' comparisons between each of GT368, GT859 and GT929 with GT136. https://www.gedmatch.com/ : accessed 20 May 2016.

[311] GEDmatch.com. Result of default 'one-to-one' comparison between GT611 and each of GT610, GT116, GT341, GT654, GT383, GT480, GT938, GT196, GT557, GT554, GT338, GT369, GT728, GT167, GT820, GT633, GT966. https://www.gedmatch.com/ : accessed 17 June 2016.

[312] FamilySearch Family Tree. *Ancestry: KW89-JLZ*. https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[313] [GT383]. *Ancestors of [GT383]*. http://freepages.folklore.rootsweb.ancestry.com/~fromheretopaternity/index.htm : accessed 1 April 2016.

[314] [GT480]. *[GT480 Family Tree]*. http://trees.ancestry.com/tree/70185236/family : accessed 1 April 2016.

[315] [GT654]. *[GT654] Family Tree*. http://trees.ancestry.com/tree/63112612/family : accessed 1 April 2016.

[316] GEDmatch.com. Result of 'Multiple Kit Analysis: Autosomal DNA comparison matrix' for GT611, GT610, GT116, GT341, GT654, GT383, GT480, GT938, GT196, GT557, GT554, GT338, GT369, GT728, GT167, GT820, GT633, GT966. https://www.gedmatch.com/ : accessed 13 May 2016.

[317] Matise Laboratory of Computational Genetics. *Map Interpolator of the Rutgers Map*. http://compgen.rutgers.edu/mapinterpolator : accessed 3 June 2016.

[318] ISOGG Wiki. *Autosomal DNA statistics*. http://www.isogg.org/wiki/Autosomal_DNA_statistics : accessed 10 December 2015.

[319] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44.

[320] GEDmatch.com. Result of ‘Multiple Kit Analysis: Autosomal DNA comparison matrix’ for GT611, GT610, GT116, GT341, GT654, GT383, GT480, GT938, GT196, GT557, GT554, GT338, GT369, GT728, GT167, GT820, GT633, GT966. https://www.gedmatch.com/ : accessed 13 May 2016.

[321] FamilySearch Family Tree. *Ancestry: KW89-JLZ.* https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[322] [GT654]. *[GT654] Family Tree*. http://trees.ancestry.com/tree/63112612/family : accessed 1 April 2016.

[323] FamilySearch Family Tree. *Ancestry: KW89-JLZ.* https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[324] [GT480]. *[GT480 Family Tree]*. http://trees.ancestry.com/tree/70185236/family : accessed 1 April 2016.

[325] FamilySearch Family Tree. *Ancestry: KW89-JLZ.* https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[326] [GT383]. *Ancestors of [GT383].* http://freepages.folklore.rootsweb.ancestry.com/~fromheretopaternity/index.htm : accessed 1 April 2016.

[327] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT999 and GT793. https://www.gedmatch.com/ : accessed 17 June 2016.

[328] GEDmatch.com. Result of default X ‘one-to-one’ comparison between GT999 and GT793. https://www.gedmatch.com/ : accessed 17 June 2016.

[329] FamilySearch Family Tree. *Ancestry: KW89-JLZ.* https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[330] [GT793]. (2016) *Re: DNA cousin on GEDmatch*. E-mail to Thad Thomas. 02:37 AM. 21 January. thad.thomas.2013@uni.strath.ac.uk.

[331] GEDmatch.com. *X ‘One-to-one’*. https://www.gedmatch.com/ : accessed 16 June 2016.

[332] Estes, Roberta. *That Unruly X….Chromosome That Is*. https://dna-explained.com/2014/01/23/that-unruly-x-chromosome-that-is/ : accessed 17 June 2016.

[333] Schaffner, Stephen F. (2004) The X chromosome in population genetics. *Nature Reviews Genetics*. 5 (1). pp. 43–51. http://www.broadinstitute.org/~sfs/nrg_Xchrom.pdf : accessed 16 June 2016.

[334] GEDmatch.com. Result of default ‘one-to-one’ comparison between PGT999M1 and GT611, GT709, GT610, GT861, GT901, GT553, GT439 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

[335] [GT999]. *[GT999] Research Tree*. http://trees.ancestry.com/tree/14578697/family : accessed 25 December 2015.

[336] [GT439]. [GT439] *Family Tree*. http://trees.ancestry.com/tree/16474504/family : accessed 25 December 2015.

[337] [GT436]. *[GT436] Family Tree*. http://person.ancestry.com/tree/243923/person/-2104873513 : accessed 27 December 2015.

[338] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT436 and GT122. https://www.gedmatch.com/ : accessed 10 June 2016.

[339] [GT439]. (2015) *Spreadsheet Template*. E-mail to Thad Thomas. 9:17 PM. 11 November. thad.thomas.2013@uni.strath.ac.uk.

[340] [GT439]. (2015) *Re: My CA with Gladys?* E-mail to Thad Thomas. 11:09 PM. 27 December. thad.thomas.2013@uni.strath.ac.uk.

[341] 23andMe, Inc. DNA Relatives [of GT999]: aggregate data. https://you.23andme.com/tools/relatives/ : accessed 10 June 2016.

[342] Matise Laboratory of Computational Genetics. *Map Interpolator of the Rutgers Map*. http://compgen.rutgers.edu/mapinterpolator : accessed 3 June 2016.

[343] [GT439]. [GT439] *Family Tree: Susan SHAW*. http://person.ancestry.com/tree/16474504/person/788502315 : accessed 12 June 2105.

[344] FamilySearch Family Tree. *Descendancy: ABIEL SHURTLEFF (LZXL-MMY)*. https://familysearch.org/tree/#view=tree&section=descendancy&person=LZXL-MMY : accessed 12 June 2105.

[345] Durand, Eric Y., Eriksson, Nicholas, & McLean, Cory Y. (2014) Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution*. 31 (8). pp. 2212–2222. http://mbe.oxfordjournals.org/content/31/8/2212 : accessed 27 November 2105.

[346] *Ibid*.

[347] ISOGG Wiki. *Autosomal DNA statistics*. http://www.isogg.org/wiki/Autosomal_DNA_statistics : accessed 10 December 2015.

[348] Bettinger, Blaine. *Visualizing Data From the Shared cM Project*. http://thegeneticgenealogist.com/2015/05/29/visualizing-data-from-the-shared-cm-project/ : accessed 16 February 2016.

[349] Speed, Doug & Balding, David J. (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*. 16 (1). pp. 33–44.

[350] GEDmatch.com. Result of ‘Multiple Kit Analysis: Autosomal DNA comparison matrix’ for GT999, GT611, GT709, GT610, GT861, GT901, GT553, GT439 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

[351] GEDmatch.com. Result of default ‘one-to-one’ comparison between GT611 and GT439. https://www.gedmatch.com/ : accessed 7 May 2016.

[352] GEDmatch.com. Result of default ‘one-to-one’ comparison between PGT999M1 and GT611, GT709, GT610, GT861, GT901, GT553, GT439 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

[353] GEDmatch.com. Result of ‘Multiple Kit Analysis: Autosomal DNA comparison matrix’ for GT999, GT611, GT709, GT610, GT861, GT901, GT553, GT439 and GT436. https://www.gedmatch.com/ : accessed 10 June 2016.

[354] [GT439]. (2015) *Spreadsheet Template*. E-mail to Thad Thomas. 9:17 PM. 11 November. thad.thomas.2013@uni.strath.ac.uk.

[355] FamilySearch Family Tree. *Ancestry: KW89-JLZ.* https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[356] [GT439]. [GT439] *Family Tree*. http://trees.ancestry.com/tree/16474504/family : accessed 25 December 2015.

[357] FamilySearch International. *Family Tree*. https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 6 December 2015.

[358] Ancestry.com. *Public Member Trees*. http://search.ancestry.com/search/db.aspx?dbid=1030 : accessed 6 December 2015.

[359] [GT999]. *Zee test Family Tree*. http://trees.ancestry.com/tree/73592017/family : accessed 25 December 2015.

[360] FamilySearch Family Tree. *Ancestry: KW89-JLZ.* https://familysearch.org/tree/#view=tree&section=fan&person=KW89-JLZ : accessed 9 June 2016.

[361] [GT436]. *[GT436] Family Tree*. http://person.ancestry.com/tree/243923/person/-2104873513 : accessed 27 December 2015.

[362] Department of Veterans Affairs. Revolutionary War Pension and Bounty-Land Warrant Application Files. SHAW, Daniel. Pension No.: W. 2447. National Archives, Washington, DC. Collection: Revolutionary War Pensions. https://www.fold3.com/image/20162877 : accessed 3 April 2016.

[363] Marriages (CR) United States. Plympton, Plymouth, Massachusetts. 6 August 1778. SHAW, Daniel and BARROWS, Mary. Collection: Massachusetts, Town and Vital Records, 1620-1988. http://search.ancestry.com/search/db.aspx?dbid=2495 : accessed 3 April 2016.

[364] Deaths (CR) United States. Rochester, Windsor, Vermont. 4 May 1804. SHAW, Mary. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1971-27949-15341-87?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > Image 74 of 368).

[365] [GT999]. *[GT999] Research Tree: Mabel [1775-]*. http://person.ancestry.com/tree/14578697/person/28455822349 : accessed 17 April 2016.

[366] Spooner’s Vermont Journal (Windsor, Vermont). (1812) List of Letters. *Spooner’s Vermont Journal*. 6 January. p. 3d; 13 January. p. 3b; 20 January. p. 4c; 27 January. p. 4c; 10 February. p. 4b. http://www.genealogybank.com/ : accessed 21 February 2016.

[367] St Lawrence Plaindealer (Canton, NY). (1937) Clarks Crossing: An Historic Spot. *St Lawrence Plaindealer*. 28 September. p. 7b&c. http://fultonhistory.com/ : accessed 2 April 2016.

[368] *Ibid*.

[369] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Elkanah. 30 November 1810. p. 27. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[370] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Salmon. 30 November 1810. p. 27, 45, 59, 83. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[371] Rogerson, Andrew E. (1858) Map of St. Lawrence Co. New York. 4/5" per mile. Philadelphia: J.B. Shields Publisher. https://www.loc.gov/item/2006626022/ : accessed 15 April 2016.

[372] Burr, David H. (1829) Map of the County of St. Lawrence. 1:151,000. New York: D H Burr. http://www.davidrumsey.com/maps4187.html : accessed 15 April 2016.

[373] Durant, Samuel W. (1878) *History of St. Lawrence Co., New York : illustrations and biographical sketches, some of its prominent men and pioneers*. Philadelphia, Pennsylvania: L.H. Everts. http://www.ancestry.com/ : accessed 16 April 2016.

[374] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Daniel, Elkanah, Daniel Jr, Salmon, Hazael, MORGAN, Forest, CHANDLER, John, and EDGERTON, Charles. 30 November 1810. p. 27, 31, 43, 45, 47, 49, 51, 59, 61, 63, 71, 83, 85, 87, 91. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[375] Jackson, Ronald V & Accelerated Indexing Systems. NY 1815 Port Arrivals. SHAW. 1815. Collection: New York, Compiled Census and Census Substitutes Index, 1790-1890. http://search.ancestry.com/search/db.aspx?dbid=3564 : accessed 3 April 2016.

[376] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 54B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[377] Census. 1860. United States. Potsdam, St Lawrence, New York. p. 68. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

[378] Marriages (CR) United States. Northfield, Washington, Vermont. 12 November 1873. MORGAN, Joseph L and HENRY, Katherine A. Collection: Vermont, Vital Records, 1720-1908. http://search.ancestry.com/search/db.aspx?dbid=4661 : accessed 3 April 2016.

[379] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. MORGAN, Forest. 30 November 1816. p. 51, 71. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[380] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. MORGAN, Forest. 30 November 1818. p. 91. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[381] Births (CR) United States. Rochester, Windsor, Vermont. 6 September 1799. SHAW, Ruel. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1971-27949-15341-87?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > Image 74 of 368).

[382] Census. 1810. United States. Rochester, Windsor, Vermont. p. 546. Collection: 1810 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7613 : accessed 3 April 2016.

[383] Census. 1850. United States. Dickinson, Franklin, New York. Schedule: Mortality. p. 1. Collection: New York, U.S. Census Mortality Schedules, 1850-1880. http://search.ancestry.com/search/db.aspx?dbid=1626 : accessed 17 April 2016.

[384] Census. 1820. United States. Potsdam, St Lawrence, New York. p. 63. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 3 April 2016.

[385] [GT999]. *Elkanah Shaw [1724-1805]*. http://person.ancestry.com/tree/14578697/person/28455768732 : accessed 17 April 2016.

[386] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 59B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[387] Marriages (CR) United States. Rochester, Windsor, Vermont. 18 March 1804. SHAW, Daniel Junr and AUSTIN, Sally. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1942-27949-15669-97?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > image 73 of 368).

[388] Green Mountain Patriot (Peacham, Vermont). (1801) Justices of the Peace. *Green Mountain Patriot*. 12 November. p. 3c. http://www.genealogybank.com/ : accessed 21 February 2016.

[389] Vermont Republican (Windsor, Vermont). (1810) Legislature of Vermont: Appointments of Justice of the Peace. *Vermont Republican*. 29 October. p. 6a. http://www.genealogybank.com/ : accessed 21 February 2016.

[390] Census. 1810. United States. Rochester, Windsor, Vermont. p. 545. Collection: 1810 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7613 : accessed 3 April 2016.

[391] Monumental inscriptions. United States. Union Cemetery, Norwood, St Lawrence, New York. 22 December 1834. SHAW, Sally. Transcribed by Anne Cady. Find A Grave Memorial: 41186042. http://www.findagrave.com/ : accessed 3 April 2016.

[392] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Daniel Jr. 30 November 1818. p. 73. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[393] Census. 1820. United States. Potsdam, St Lawrence, New York. p. 63. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 3 April 2016.

[394] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 5B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[395] Notice of probate. (1846) *St. Lawrence Republican*. CHANDLER, John. 28 July. p. 4d. http://nyshistoricnewspapers.org/lccn/sn83031401/1846-07-28/ed-1/seq-4 : accessed 15 April 2016.

[396] Monumental inscriptions. United States. Bayside Cemetery, Potsdam, St Lawrence, New York. 10 December 1860. CHANDLER, Waitstill Shaw. Transcribed by gravehunter1218. Find A Grave Memorial: 75463137. http://www.findagrave.com/ : accessed 7 April 2016.

[397] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. CHANDLER, John. 30 October 1811. p. 31, 47, 61, 85. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[398] Census. 1820. United States. Potsdam, St Lawrence, New York. p. 62. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 7 April 2016.

[399] Census. 1830. United States. Potsdam, St Lawrence, New York. p. 135. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058 : accessed 7 April 2016.

[400] Census. 1840. United States. Potsdam, St Lawrence, New York. p. 199. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

[401] Census. 1860. United States. Potsdam, St Lawrence, New York. p. 781. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

[402] Monumental inscriptions. United States. Lakenham Cemetery, Carver, Plymouth, Massachusetts. 18 June 1781. SHAW, Waitstill. Transcribed by Anne Shurtleff Stevens. Find A Grave Memorial: 36893709. http://www.findagrave.com/ : accessed 7 April 2016.

[403] Deaths (CR) United States. Carver, Plymouth, Massachusetts. 18 June 1781. SHAW, Waitstill. Collection: Massachusetts, Town and Vital Records, 1620-1988. http://search.ancestry.com/search/db.aspx?dbid=2495 : accessed 7 April 2016.

[404] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 56A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[405] Potsdam Herald-Recorder. (1934) Speaks on ‘The Union’. *The Potsdam Herald-Recorder.* 14 September. p. 1e, 3c&d. http://nyshistoricnewspapers.org/lccn/sn84035824/1934-09-14/ed-1/seq-1 : accessed 17 April 2016.

[406] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 56A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[407] Census. 1860. United States. Potsdam, St Lawrence, New York. p. 820. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

[408] Census. 1820. United States. Potsdam, St Lawrence, New York. p. 63. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 3 April 2016.

[409] Census. 1830. United States. Potsdam, St Lawrence, New York. p. 134. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058 : accessed 7 April 2016.

[410] Census. 1840. United States. Potsdam, St Lawrence, New York. p. 209. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

[411] Census. 1860. United States. Potsdam, St Lawrence, New York. p. 820. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

[412] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 56A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[413] Ancestry.com. Town Clerks´ Registers of Men Who Served in the Civil War, ca 1861-1865. SHAW, Elkanah. Collection: New York, Town Clerks’ Registers of Men Who Served in the Civil War, ca 1861-1865. http://search.ancestry.com/search/db.aspx?dbid=1964 : accessed 8 April 2016.

[414] [GT999]. *Elkanah Shaw [1724-1805]*. http://person.ancestry.com/tree/14578697/person/28455768732 : accessed 17 April 2016.

[415] Adjutant General’s Office. Register of Enlistments in the U.S. Army, 1798-1914. SHAW, Hazael. 21 August 1812. Record Group 94: Records of the Adjutant General’s Office, 1762 - 1984. Microfilm Publication M233. National Archives, Washington, DC. Collection: U.S. Army, Register of Enlistments, 1798-1914. http://search.ancestry.com/search/db.aspx?dbid=1198 : accessed 15 April 2016.

[416] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Hazael. 23 August 1817. p. 63, 87. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[417] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. SHAW, Hazael. 23 August 1817. p. 63, 87. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[418] Death announcements. (1895) *Ogdensburg advance and St. Lawrence weekly Democrat*. MAYON[sic], Cynthia. 28 February. p. 1b. http://nyshistoricnewspapers.org/lccn/sn83031423/1895-02-28/ed-1/seq-1/ : accessed 20 April 2016.

[419] Testamentary records. United States. 23 February 1893. SHAW, Daniel. Notice of Probate. Collection: New York, Wills and Probate Records, 1659-1999. http://search.ancestry.com/search/db.aspx?dbid=8800 : accessed 20 April 2016.

[420] Deaths (CR) United States. Rochester, Windsor, Vermont. 10 August 1803. SHAW, Freeman. Source film no.: 28744. Collection: Vermont, Town Clerk, Vital and Town Records, 1732-2005. https://familysearch.org/pal:/MM9.3.1/TH-1971-27949-15341-87?cc=1987653 : accessed 3 April 2016. (Windsor > Rochester > Town records 1782-1860 > Image 74 of 368).

[421] Monumental inscriptions. United States. Union Cemetery, Norwood, St Lawrence, New York. 20 December 1865. EDGERTON, Elizabeth Shaw. Transcribed by Anne Cady. Find A Grave Memorial: 41185980. http://www.findagrave.com/ : accessed 3 April 2016.

[422] Census. 1860. United States. Potsdam, St Lawrence, New York. p. 820. Collection: 1860 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7667 : accessed 3 April 2016.

[423] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 54A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 7 April 2016.

[424] Testamentary records. United States. 4 January 1850. EDGERTON, Charles. Administration. Collection: New York, Wills and Probate Records, 1659-1999. http://search.ancestry.com/search/db.aspx?dbid=8800 : accessed 17 April 2016.

[425] Testamentary records. United States. 13 December 1855. EDGERTON, Ransom G. Administration. Collection: New York, Wills and Probate Records, 1659-1999. http://search.ancestry.com/search/db.aspx?dbid=8800 : accessed 17 April 2016.

[426] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. EDGERTON, Charles. 19 September 1812. p. 43. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[427] Raymond, Benjamin. Benjamin Raymond Record of Sales 1803-1818. EDGERTON, Charles. 30 November 1816. p. 71. New York Heritage Digital Collections: Potsdam Public Museum. 1954-176-2.3. http://cdm16694.contentdm.oclc.org/cdm/compoundobject/collection/ppm/id/529/ : accessed 15 April 2016.

[428] Census. 1820. United States. Potsdam, St Lawrence, New York. p. 62. Collection: 1820 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7734 : accessed 7 April 2016.

[429] Census. 1830. United States. Potsdam, St Lawrence, New York. p. 135. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058 : accessed 7 April 2016.

[430] Census. 1840. United States. Potsdam, St Lawrence, New York. p. 199. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

[431] Census. 1850. United States. Potsdam, St Lawrence, New York. Schedule: Agriculture. p. 531. Collection: Selected U.S. Federal Census Non-Population Schedules, 1850-1880. http://search.ancestry.com/search/db.aspx?dbid=1276 : accessed 7 April 2016.

[432] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 54A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 7 April 2016.

[433] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 10A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[434] Census. 1900. United States. Tracy, Lyon, Minnesota. p. 10B. Collection: 1900 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7602 : accessed 21 April 2016.

[435] Census. 1900. United States. Potsdam, St Lawrence, New York. p. 10A. Collection: 1900 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7602 : accessed 21 April 2016.

[436] Deaths (CR) United States. Winona, Winona, Minnesota. 24 February 1913. CLARK, Eliza Rebecca. Source film no.: 2138535. Cert no.: 15696. Collection: Minnesota Deaths and Burials, 1835-1990. https://familysearch.org/ark:/61903/1:1:FD9Y-J9W : accessed 3 April 2016.

[437] Monumental inscriptions. United States. Bayside Cemetery, Potsdam, St Lawrence, New York. 18 March 1855. TRAVER, Susan. Transcribed by gravehunter1218. Find A Grave Memorial: 148278717. http://www.findagrave.com/ : accessed 3 April 2016.

[438] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 10A. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[439] Census. 1900. United States. Tracy, Lyon, Minnesota. p. 10B. Collection: 1900 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7602 : accessed 21 April 2016.

[440] Census. 1900. United States. Potsdam, St Lawrence, New York. p. 10A. Collection: 1900 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=7602 : accessed 21 April 2016.

[441] Frisbie, Richard. *Vermont/New York Boundary History*. http://www.hopefarm.com/vermont.htm : accessed 21 April 2016.

[442] Census. 1830. United States. Potsdam, St Lawrence, New York. p. 151. Collection: 1830 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8058 : accessed 7 April 2016.

[443] Census. 1840. United States. Potsdam, St Lawrence, New York. p. 214. Collection: 1840 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8057 : accessed 7 April 2016.

[444] Census. 1850. United States. Potsdam, St Lawrence, New York. p. 5B. Collection: 1850 United States Federal Census. http://search.ancestry.com/search/db.aspx?dbid=8054 : accessed 3 April 2016.

[445] Jackson, Ronald V & Accelerated Indexing Systems. NY 1815 Port Arrivals. SHAW. 1815. Collection: New York, Compiled Census and Census Substitutes Index, 1790-1890. http://search.ancestry.com/search/db.aspx?dbid=3564 : accessed 3 April 2016.

[446] Monumental inscriptions. United States. North Bridgewater Cemetery, Bridgewater, Windsor, Vermont. 15 August 1850. SHAW, Elkanah. Transcribed by David Edsall. Find A Grave Memorial: 111849132. http://www.findagrave.com/ : accessed 17 April 2016.

[447] [GT999]. *Elkanah Shaw [of Potsdam, St Lawrence, New York]*. http://person.ancestry.com/tree/14578697/person/28455833090 : accessed 17 April 2016.

[448] [GT999]. *Elkanah Shaw [of Bridgewater, Windsor, Vermont]*. http://person.ancestry.com/tree/14578697/person/67001106167 : accessed 17 April 2016.

[449] FamilySearch International. *GEDCOM X and the Genealogical Research Process*. http://www.gedcomx.org/GEDCOM-X-and-the-Genealogical-Research-Process.html : accessed 21 April 2016.

Journal of Genetic Genealogy

AUTOSOMAL MATCH IN-COMMON-WITH CLUSTER ANALYSIS WITH NETWORK VISUALIZATION TOOLS: AN EXAMPLE USING THE GEPHI OPEN-SOURCE TOOL

By J. David Vance

Abstract

The genetic genealogy community has many tools for autosomal DNA analysis, and many tools and techniques have been developed to use autosomal DNA match inter-relationships to assist in the identification of common ancestors. Many of these techniques work best with matches who share larger amounts of DNA and are therefore closer relatives whose genealogical connections are more readily discovered. This review discusses the merits of yet another technique, network visualization, which can cluster large datasets of matches even lower than 20 cMs (in this review, down to 7 cMs) and can identify and analyze clusters of In-Common-With matches which, especially when combined with other genealogical information like known relationships of certain matches or clusters determined by other methods, can help focus and prioritize our analysis of matches to find our shared ancestry and thereby extend our genealogical knowledge. In this review the Gephi tool was used as the network visualization platform but the approach is independent of the specific tool.

Background Assumptions

This review is not intended for those new to autosomal DNA analysis; not because the techniques are difficult to understand but because there are more commonly-suggested starting analysis techniques like the Leeds Method or even the analysis tools provided by commercial companies, and this review covers an approach which might be more useful after those starting techniques have been exhausted.

For that reason, this review does assume that readers have a basic familiarity with other autosomal DNA match analysis techniques like the Leeds Method and some fundamentals of autosomal DNA analysis for genealogy like the relationship of shared autosomal DNA segments to genealogical relationships between matches.

Introduction

When many of us take our first autosomal DNA (atDNA) test what we are hoping to find are matches who will help us figure out the gaps in our knowledge of our own ancestry. We hope for as large a pool of matches as possible with the somewhat mixed blessing that we then have to untangle the often difficult questions of how they all may be related to us and to each other.

Very often a key subset of matches will share larger centimorgans (cMs) of DNA with us and our relationship with those matches will be closer and clearer (say, perhaps within 3rd cousins). For this subset, where the genealogical relationships between ourselves and those matches don't become clear by simply comparing our known genealogies we can often ferret out the relationships using our better-known atDNA analysis techniques: Leeds Method, segment matching, and so on, or using the suite of deservedly-popular tools which have been developed by the commercial companies and third parties in support of those techniques.

The success rate of our most common techniques though drops off rapidly with more distant relationships and as the shared DNA segments get smaller. While sometimes there is no substitute for doggedly researching and comparing genealogies to find common ancestors, these more distant matches can still be frustrating for genetic genealogists especially if there are a large number of matches in the “4th cousin and further” category whose genealogical relationships are unclear and who are especially resistant to analysis with our most common techniques.

A lesser-known approach especially for tackling these more distant matches is analysis using network visualization software to group them into In-Common-With (ICW) clusters – groups of matches who are themselves matches to each other and who may as a result all descend from a common ancestor.

This is not a new approach – network visualization approaches have been used for ICW cluster analysis at least as far back as 2017 by Barbara Griffiths and Shelley Crawford using a variety of tools including Pajek and NodeXL. In this review we have used the Gephi tool (free and open-source at https://gephi.org/) as the clustering and graphing platform but except for different flexibility in clustering and filtering options, the approach is independent of the tool and any similar network visualization software package would support the same approach.

Others are also applying network visualization to their own autosomal data analysis using similar but not identical approaches to the examples shown in this review. Their results are often displayed on social media forums to other genetic genealogists and generate much surprise and discussion, which suggests that the techniques are not widely practiced and might benefit more people if they were more widely understood.

Network visualization is not a replacement for more common analysis techniques; in fact as a clustering approach for ICW matches it is very similar to a Leeds Method analysis though more complicated to set up the necessary data and to analyze. This is one reason that a simpler method like Leeds should be attempted first, but another reason is that as we will explain in this review, mapping an initial subgroup of matches to their genealogical relationships using other methods first can be extended by network visualization to wider clusters of more distant matches. This means that network visualization is not only a stand-alone technique but also an approach that can extend the results of prior analysis.

The other advantage of network visualization is that it can be used to very quickly sort large networks into clusters and then explore these clusters to investigate their shared origins, and to analyze the network as a whole through the application of different filters that highlight important relationships both within and across clusters.

While we present a few examples of network visualization analysis of ICW clusters in this review, we would also propose that network visualization should be considered a more general approach which may include several analysis techniques depending on desired outcome and number and type of autosomal matches. Our main point in writing this review is simply to show by example the merits of network visualization as an important tool in an analysis toolkit, not to suggest that there is one best network visualization analysis approach.

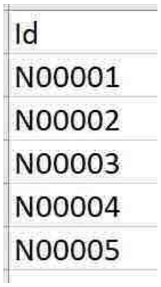
In this review we show by example how a network visualization tool like Gephi can be used to sort a large ICW network into clusters which include matches at smaller levels of shared cMs. We offer one approach for extending the relationship knowledge of a small number of matches out to the rest of the identified clusters, and also show how the tool's filtering options can be used to dissect the network in different ways to gain additional insights.

Methods and Data

Preparing the Network Data

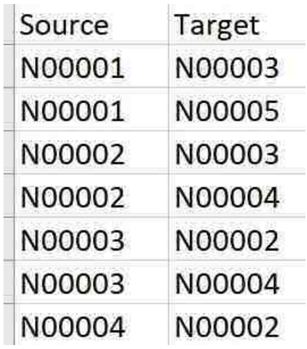
A network for Gephi purposes is simply a set of "nodes", each connected to other nodes via "edges" which are represented by connecting lines. For the tool to represent a network the minimum required data therefore are a "node table" which lists all the network's nodes, and an "edge table" which has pairs of node ids representing the two endpoints of each line in the network.

At a minimum then this could be represented by two tables, a Node Table as in the example in Figure 1, and an Edge Table as in the example in Figure 2.



Id
N00001
N00002
N00003
N00004
N00005

Figure 1a. Simple Node Table



Source	Target
N00001	N00003
N00001	N00005
N00002	N00003
N00002	N00004
N00003	N00002
N00003	N00004
N00004	N00002

Figure 2. Simple Edge Table

In this example we will use an undirected network, so which node is "Source" versus "Target" is immaterial (and note that while there are duplicate connections shown in the edge table in Figure 2, Gephi removes those on undirected networks). More complex network analyses however could be conducted using directed networks or even weighted edges (for example weighted by number or size of shared cM segments, etc).

To build this network we will use the match lists provided by an autosomal DNA testing company. Most of the major testing companies (Ancestry, Family Tree DNA, and 23andMe as examples) provide this level of detail, although not all of them provide it in easily downloadable files.

The Node Table of course is simply our list of autosomal matches, while the Edge Table lists which of our matches also match each other. These lists can be built by hand, though the commercial companies do not all provide easy identification of matches that also match each other.

A more automated method of producing these files is to use the DNAGedcom tool; their "match" and "icw" (in common with) output files can be used for Node and Edge tables, respectively, for this purpose and only columns headers need to be changed (Gephi requires "Id" as the header in the Node Table and “Source” and “Target” as the headers in the Edge Table). At the time of this publication not all commercial companies however allow the use of DNAGedcom to download their data, so this should be investigated before attempting that method. Or ICW information can also be obtained from an automated cluster assessment from DNA Painter or Genetic Affairs.

Brit Nicholson has an excellent tutorial on building these files from GedMatch data in his Aug 2020 blog post on https://www.dna-sci.com.

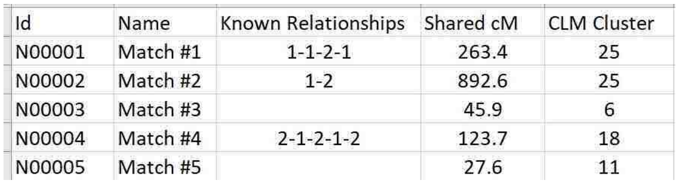
Otherwise these files can also be built manually using the reports on the testing companies’ websites (For example the Shared Matches report for AncestryDNA, or the Relatives in Common report from 23andMe).

For the example in this review, DNAGedcom was used to create the original ICW data files.

One powerful addition to the network data is the ability to include additional columns in the Node Table and use them for additional analysis. For our actual example in this review, we will add four additional columns:

1. Match names, which normally would be the given names of matches but will here be represented by the labels “Match #1”, “Match #2”, etc.;
2. Known Genealogical Relationships, discussed in the next section;
3. Shared cM between that match and our DNA, so that the network can include this for filtering and analysis purposes;
4. The cluster numbers resulting from a Collins-Leeds Method analysis of the matches with larger shared cMs; this was done only for test purposes to show how network visualization compares to Leeds Method approaches and is not otherwise necessary for network analysis.

These additional columns are shown in Figure 3.



Id	Name	Known Relationships	Shared cM	CLM Cluster
N00001	Match #1	1-1-2-1	263.4	25
N00002	Match #2	1-2	892.6	25
N00003	Match #3		45.9	6
N00004	Match #4	2-1-2-1-2	123.7	18
N00005	Match #5		27.6	11

Figure 3. Adding More Columns to the Node Table.

The “CLM Cluster” column used in this example was added solely for this review and would not normally be required. The Collins-Leeds Method was run on the same match data for comparison to network visualization; as shown later in our example, clusters using Gephi match up with clusters identified through Leeds Method analysis but can be more easily filtered to visualize and highlight relationships.

To populate this column, the Collins-Leeds Method identified clusters for some 2,870 matches and these cluster numbers were listed for those matches in a new column of the Nodes Table as shown in Figure 3. This data will be used later in this review as a comparison between clustering methods.

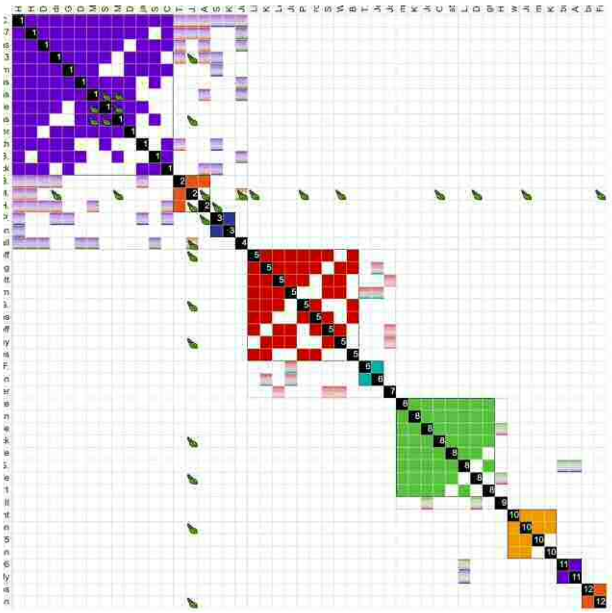


Figure 4. Collins-Leeds Method run on the ICW data for comparison purposes. This was used only to identify common clusters for a subset of ICW matches (names have been omitted for privacy).

Representing Known Genealogical Relationships in the Data

Usually with our match lists there is at least a small subset where we already know the genealogical relationships between us and those matches. That subset may include known relatives, matches with whom we have compared genealogies and identified the common ancestors, or even matches whose genealogy we have built out through traditional research methods and have been successful in identifying the connections.

If this subset exists, the known genealogical connections can be helpful in identifying the common ancestors of clusters and, by extension, the common ancestors of other matches to each other and with ourselves.

To demonstrate this in this example, we are using a modified Ahnentafel numbering system, where male ancestors are represented by “1” and female ancestors represented by “2”, and each generation is represented. Therefore our father is “1”, mother is “2”, father’s father is “1-1”, father’s mother is “1-2”, and so on. Figure 5 also includes an example match on the left who descends from our great-grandparents on our maternal grandfather’s side. Since both we and the match inherited DNA from the pair of ancestors marked “2-1-1” and “2-1-2”, we represent this as “2-1-1/2” in the Known Relationships column.

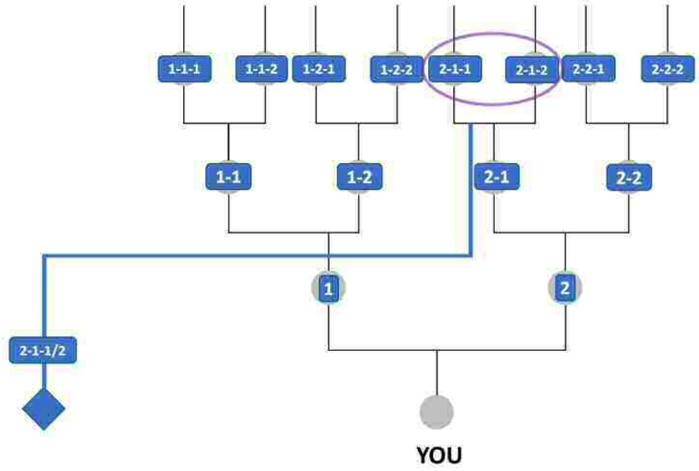


Figure 5. Keeping track of Known Relationships

The purpose of this column will become more clear in the example below, but this nomenclature was selected because it identifies the shared ancestral lines and number of generations between us and certain matches.

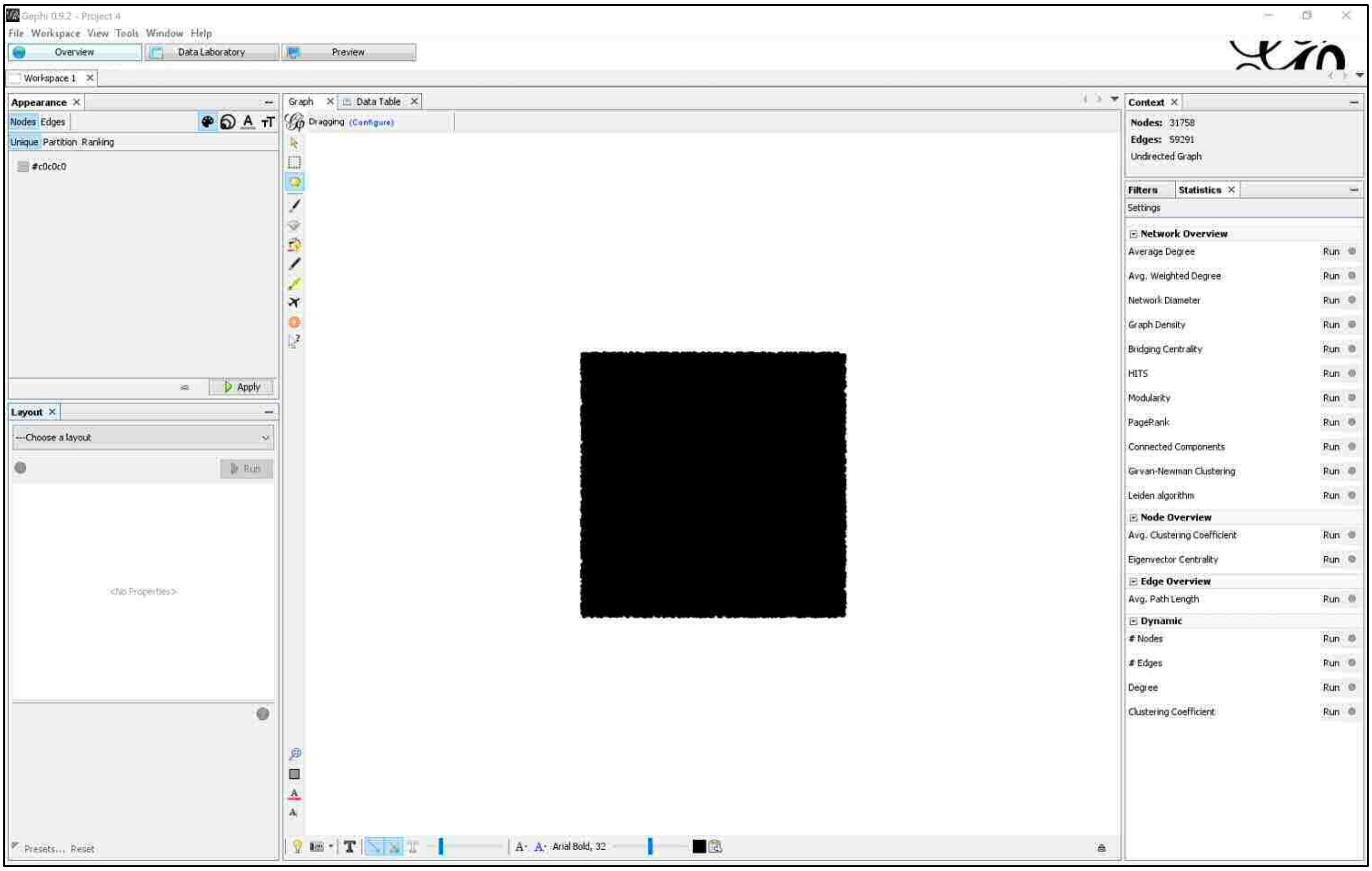
In our example, we have identified known genealogical relationships for approximately 200 of the closer matches and this information has been captured using this custom nomenclature.

Combining these methods for one of the author’s autosomal DNA tests resulted in two files to use as input for our example: a Nodes file of 31,758 matches with extra columns as shown in Figure 3, with some 200 of these matches marked with Known Genealogical Relationships and 2,870 of the matches marked with a CLM Cluster number from the Collins-Leeds Method clustering mentioned earlier.

Creating the Network

Installation of the Gephi tool is beyond the scope of this review, but requires the correct Java runtime environment and enough CPU to handle the intensive computational requirements. This example was created using Gephi version 0.9.2 on a Microsoft Surface 3 running Windows 10 with an i7-1065G7 CPU and 3GB of RAM.

The tool allows the import of data as a Node Table followed by the Edge Table; and when the data is first imported it is displayed in the Graph display and initially appears as a square without identifiable structure or color, as shown in Figure 6.



Nodes: 31758
Edges: 59291
Undirected Graph

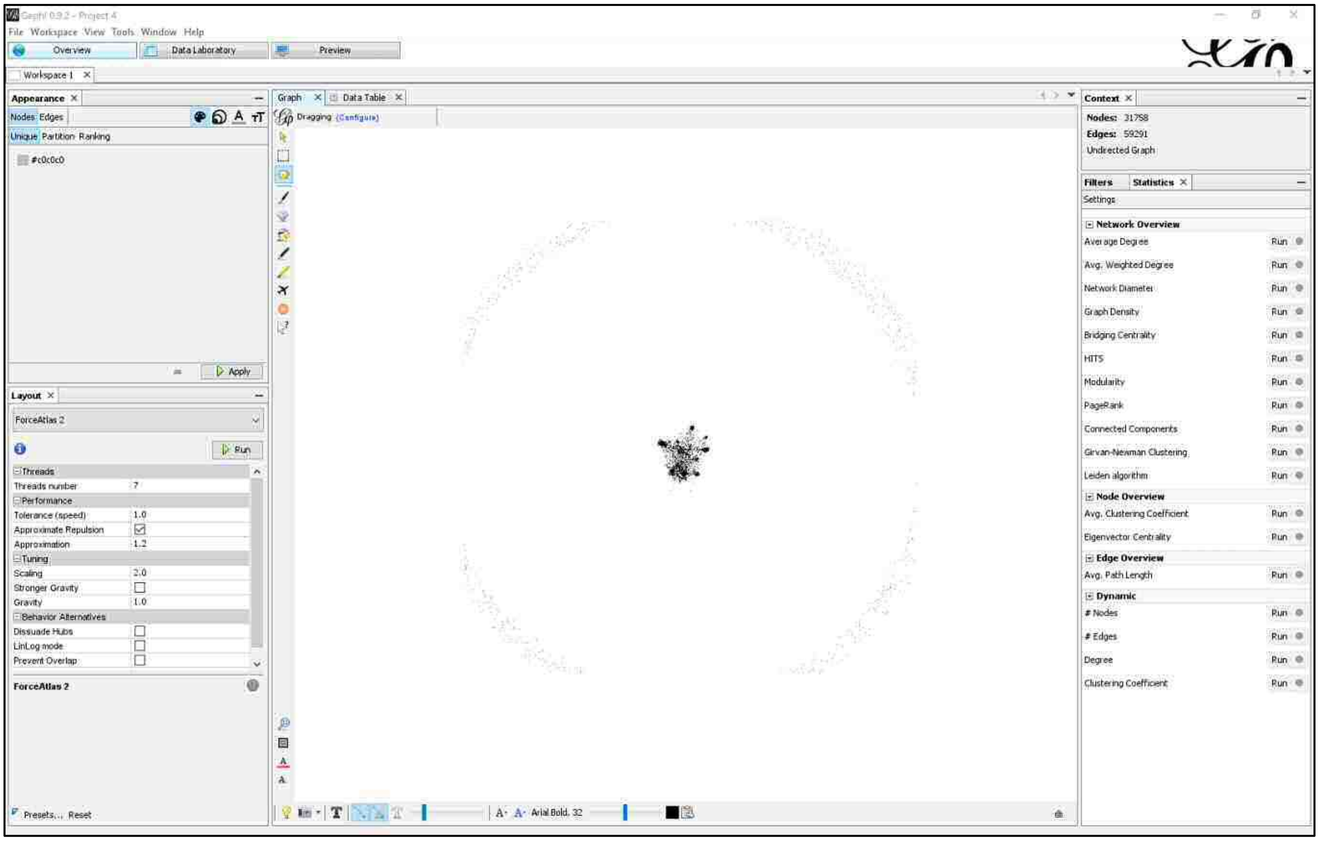
Figure 6. What a network in Gephi first looks like.

A complete overview of Gephi functions is also outside the scope of this review; in brief the "Appearance" window in the upper left controls the appearance of the graphical network display – colors, size of nodes, etc; while the "Layout" window in the lower left controls the clustering analysis methods that can be applied to the network. In the upper right the "Context" menu gives information about the current network or portion thereof which is currently being displayed, while the "Statistics" menu in the lower right allows various statistics about the network to be calculated. The "Filters" tab to the left of the "Statistics" tab further allows the flexible application of a number of filters. These options will all be important in this example.

Applying a Layout

In the lower left window, Gephi will show various layouts that can be applied to the network to arrange the nodes in the 3D display graph according to various criteria. Which one to use depends on the characteristics of the network, though for networks of ICW matches the "Force Atlas 2" layout appears to work best. In simple terms, this spatialization algorithm causes nodes to "repulse" each other and higher numbers of edges between nodes to be "attracted" to each other and as the algorithm is run the network visualization stretches in real-time into clusters of nodes which share higher numbers of edges with each other. This algorithm was developed by the authors of Gephi and is more fully described in this published study: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679.

After applying the Force Atlas 2 layout (and after zooming the graph display out to include all nodes) the graph looks like Figure 7.



ForceAtlas 2

Threads	
Threads number	7
Performance	
Tolerance (speed)	1.0
Approximate Repulsion	☑
Approximation	1.2
Tuning	
Scaling	2.0
Stronger Gravity	☐
Gravity	1.0
Behavior Alternatives	
Dissuade Hubs	☐
LinLog mode	☐
Prevent Overlap	☐

Nodes: 31758
Edges: 59291
Undirected Graph

Network Overview	
Average Degree	Run
Avg. Weighted Degree	Run
Network Diameter	Run
Graph Density	Run
Bridging Centrality	Run
HITS	Run
Modularity	Run
PageRank	Run
Connected Components	Run
Girvan-Newman Clustering	Run
Leiden algorithm	Run
Node Overview	
Avg. Clustering Coefficient	Run
Eigenvector Centrality	Run
Edge Overview	
Avg. Path Length	Run
Dynamic	
# Nodes	Run
# Edges	Run
Degree	Run
Clustering Coefficient	Run

Figure 7. Graph after Force Atlas 2 layout applied.

The circle of nodes flung out to the edges of the graph represent the matches in the data set who are only matches with the autosomal DNA test used and have no other matches to anyone else – in other words, they are not "In Common With" matches. In the network representation, they have no edges with any other nodes and so are "repulsed" to the far outer edges of the graph. While individually they may certainly be matches worth pursuing especially if some share larger segments of DNA with the autosomal test used, this example will ignore them to focus on the clustered network at the center, which represents the matches which do have one or more matches among the others in the data.

Discussion

Characteristics of Network Clusters

At the center of this network graph, the nodes which do share edges with each other form into clusters, analogous to the squares in a Leeds Method analysis (this will be demonstrated shortly).

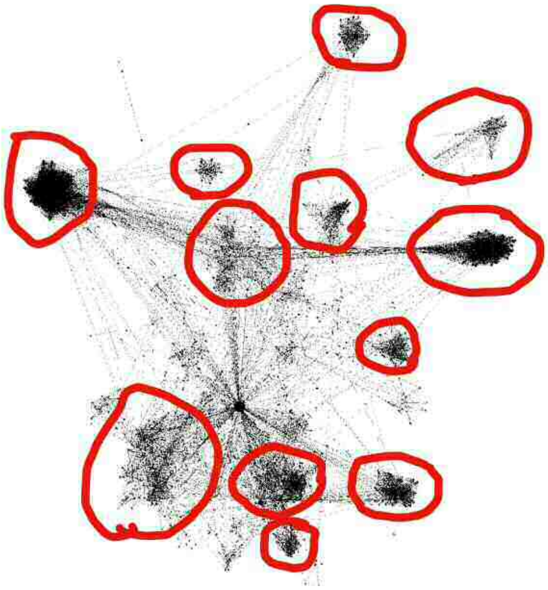


Figure 8. Nodes form clusters of various sizes and density.

Just as with a Leeds Method analysis, these clusters can be matched with groups who inherited common DNA segments from common ancestors. Clusters will vary in size and density according to the number of ICW matches, and because the Force Atlas 2 visualization algorithm causes nodes with higher interconnectivity to be attracted more strongly to each other, denser clusters represent subgroups of the network who have a higher incidence of ICW matching between each other than with the rest of the network. By definition, a subgroup of the network with a higher incidence of ICW matching will correspond to the descendants of common shared ancestors who share DNA from that same origin; however these common ancestors may be from a wide range of generations back in time, and some members of a cluster may be related more closely by later common ancestors who passed the same DNA along. So the shape and distribution of clusters will be very dependent on the closeness and degree of interrelationships between ICW matches for any given autosomal DNA tester.

Just as for Leeds Methods, clusters will also be less well-defined due to several influences, including:

1. Endogamy and/or pedigree collapse, which will cause clustering of matches who do share DNA from common ancestors but whose relationships will be from multiple common ancestral paths and not easily assignable to one common origin;
2. Pile-up regions of individual chromosomes, which will cause unrelated matches to have perceived ICW relationships that will show as connections between clusters on the graph, and may in significant cases blur the boundaries of otherwise unrelated clusters;
3. Very close relatives to the test being analyzed, who share DNA from several common lines of descent with the given autosomal tester and whose nodes will therefore show connections across many clusters.

All three of these influences can be somewhat mitigated by deeper analysis of the matches, shared DNA segments, and relationships, and changes to the network data to eliminate their influences. In particular, the third influence (close relatives) may be the easiest to identify and mitigate.

Changing the node sizes (in the Appearance Window) to reflect the amount of shared cMs, we can easily identify close relatives by node size. Figure 9 for example shows a portion of the network visualization where a close relative (a full sibling, in this case) has connections across various clusters.

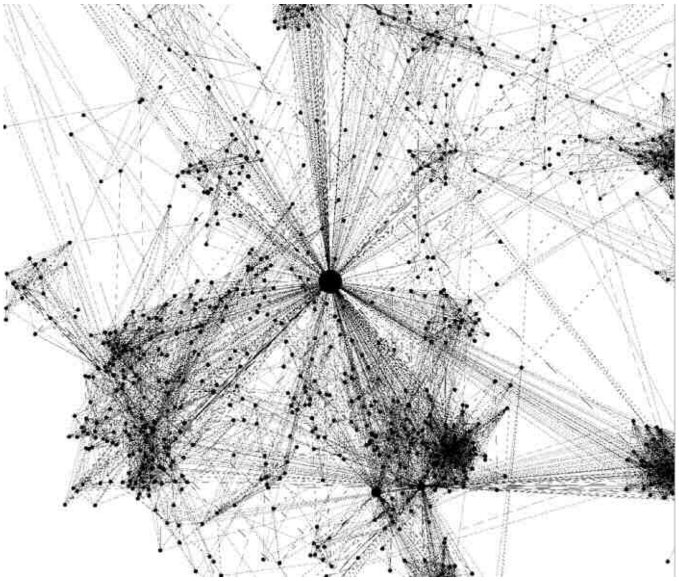


Figure 9. A subset of Figure 8 showing a full sibling crossing many clusters.

In some cases close relatives may be useful to retain; first cousins for example will generally show connections only to father’s-side or mother’s-side clusters which would aid the identification of origins. But for analysis of more distant clusters, their influence obscures the clear identification of clusters and close relatives can also simply be removed from the data set. For example, removing this full sibling from the data (and reapplying the Force Atlas 2 layout) would give us Figure 10, where cluster definition has significantly improved.

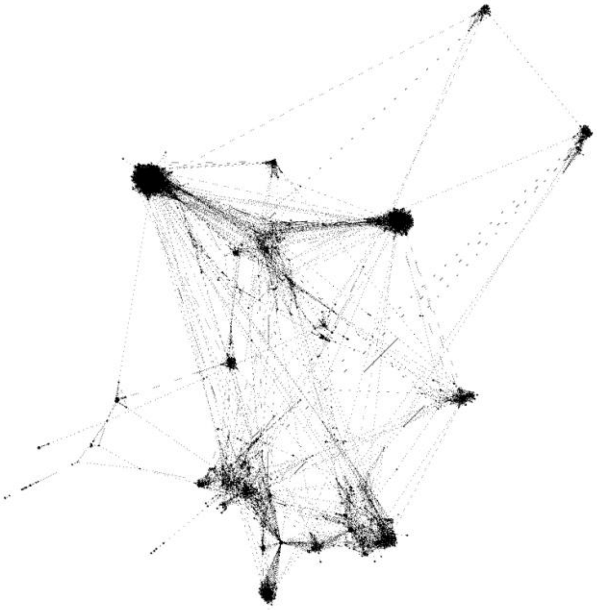
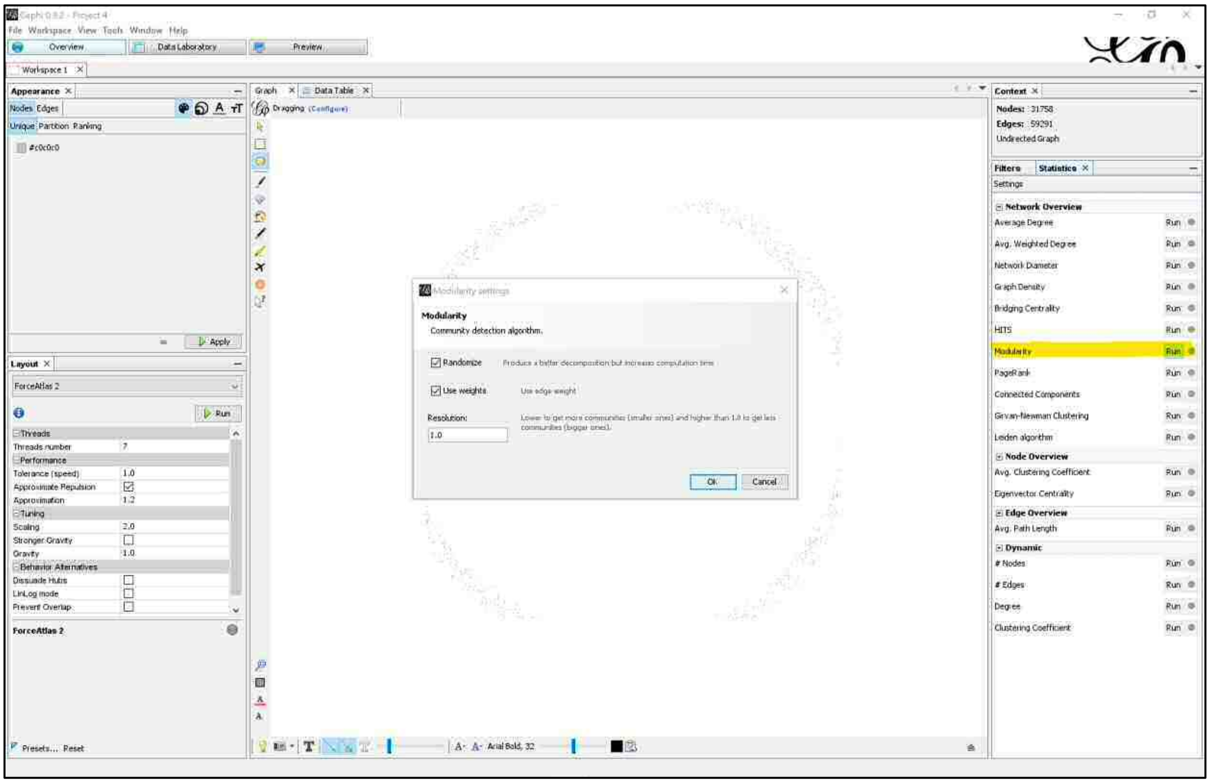


Figure 10. The network without the full sibling node.

Applying Color

The measure of the strength of division of a network into clusters is called *modularity*. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Gephi offers a calculation of modularity among its statistics functions in the right-hand window. The advantage of running these calculations is that they also create sub-components of the network which can then be colored in the Appearance window.



#c0c0c0

ForceAtlas 2

Threads	
Threads number	7
Performance	
Tolerance (speed)	1.0
Approximate Repulsion	☑
Approximation	1.2
Tuning	
Scaling	2.0
Stronger Gravity	□
Gravity	1.0
Behavior Alternatives	
Dissuade Hubs	□
LinLog mode	□
Prevent Overlap	□

Modularity settings

Modularity

Community detection algorithm.

☑ Randomize — Produce a better decomposition but increases computation time

☑ Use weights — Use edge weight

Resolution: 1.0 — Lower to get more communities (smaller ones) and higher than 1.0 to get less communities (bigger ones).

OK Cancel

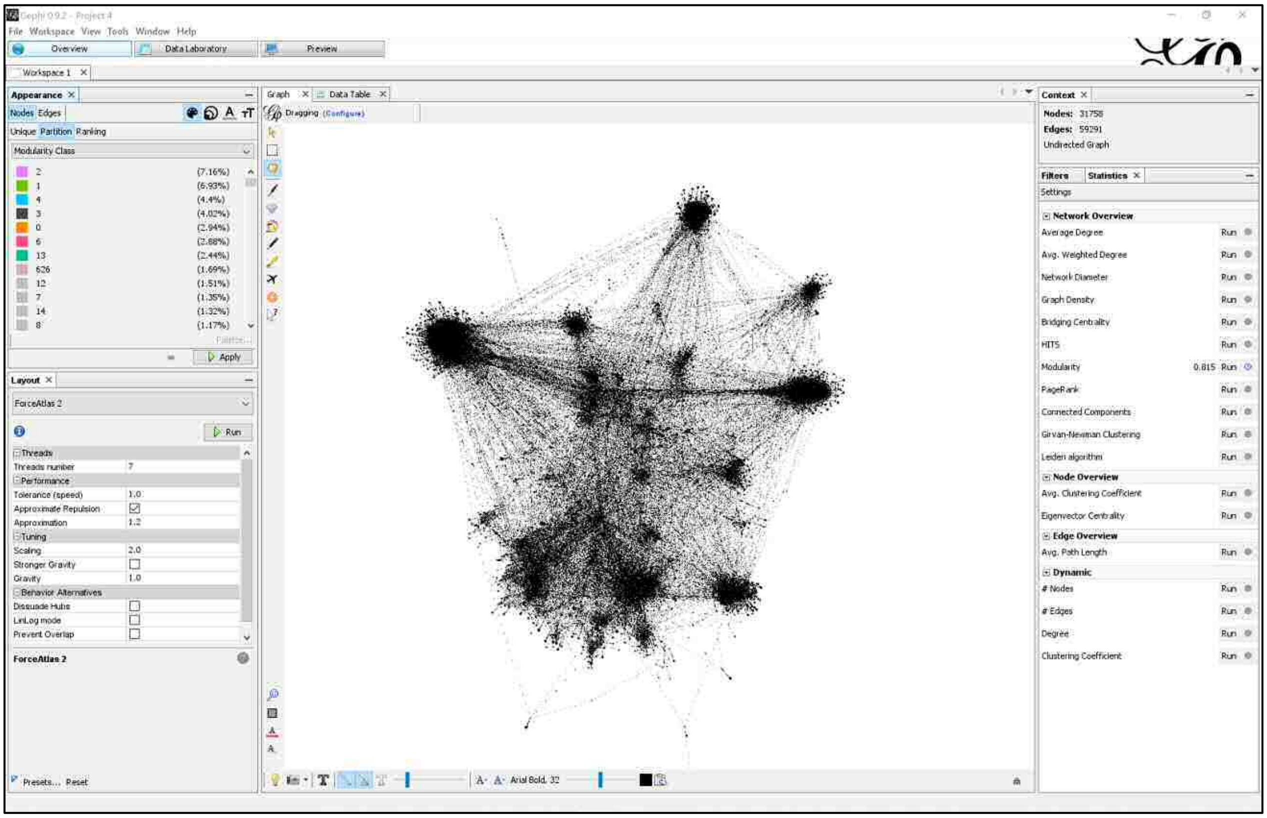
Nodes: 31758
Edges: 59291
Undirected Graph

Network Overview	
Average Degree	Run
Avg. Weighted Degree	Run
Network Diameter	Run
Graph Density	Run
Bridging Centrality	Run
HITS	Run
Modularity	Run
PageRank	Run
Connected Components	Run
Girvan-Newman Clustering	Run
Leiden algorithm	Run
Node Overview	
Avg. Clustering Coefficient	Run
Eigenvector Centrality	Run
Edge Overview	
Avg. Path Length	Run
Dynamic	
# Nodes	Run
# Edges	Run
Degree	Run
Clustering Coefficient	Run

Figure 11. Running the Modularity Statistic

For the analysis of ICW matches, the "Randomize" option can be left checked although the "Use Weights" option is irrelevant since the edges are not weighted (although an extension of this approach might find it useful to weight them by shared cM or other additional information). The function also requires a "Resolution" number which drives the number of sub-networks that the modularity will determine. In order to color the visible clusters on the network, different resolution numbers may have to be used until the color groups match up as well as possible to the visible clusters. For this purpose a resolution of 3.0 seemed to work well for this data set, but it will vary somewhat according to the specific set of ICW matches.

Once the modularity function has completed it will produce a report, which for our purposes can be ignored. More usefully it will also have added a new column in the data called Modularity Class which labels the nodes by the divisions found by the modularity algorithm. In the Appearance window on the left, nodes can be colored by this Modularity Class. Before-and-after graphs of the central network of this coloring are shown in Figures 12 and 13.



Modularity Class

🟪	2	(7.16%)
🟩	1	(6.93%)
🟦	4	(4.4%)
⬛	3	(4.02%)
🟧	0	(2.94%)
🟥	6	(2.88%)
🟩	13	(2.44%)
▨	626	(1.69%)
▨	12	(1.51%)
▨	7	(1.35%)
▨	14	(1.32%)
▨	8	(1.17%)

ForceAtlas 2

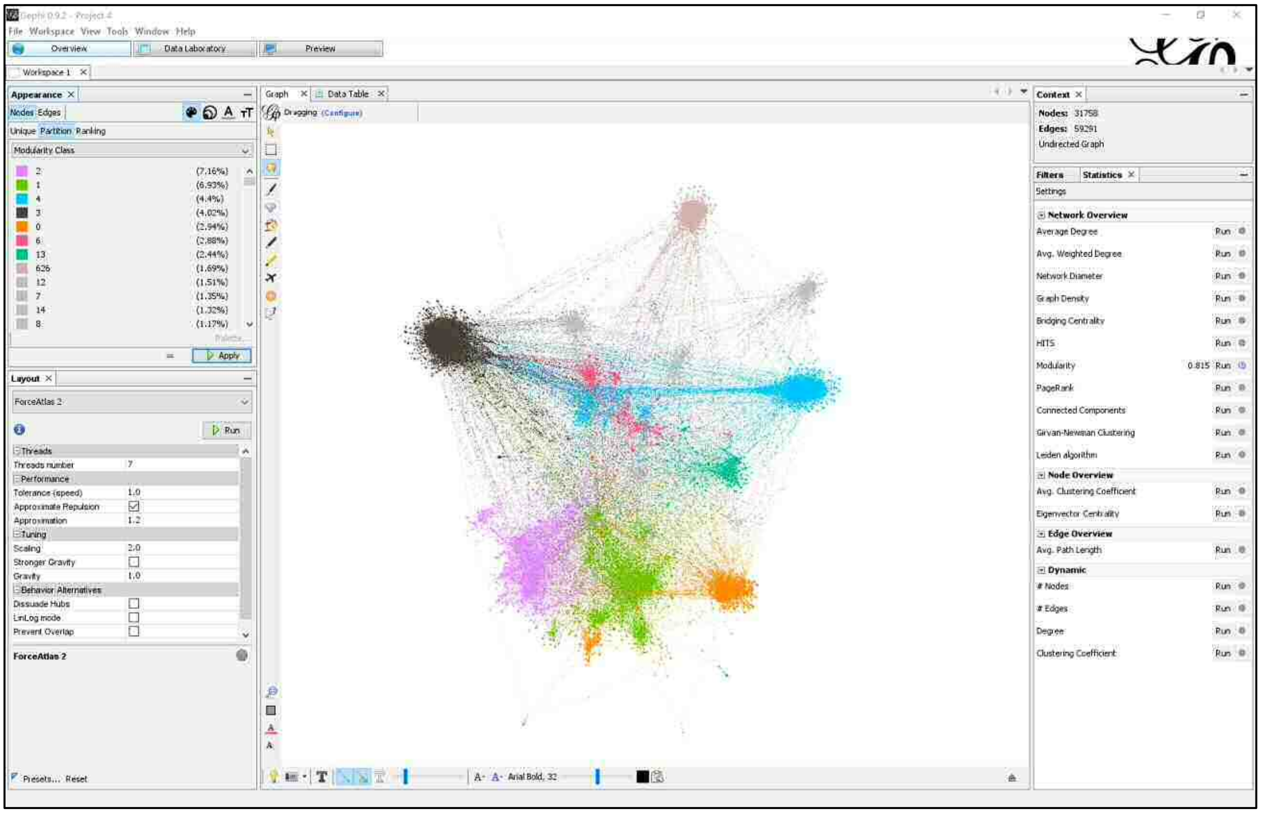
Threads	
Threads number	7
Performance	
Tolerance (speed)	1.0
Approximate Repulsion	☑
Approximation	1.2
Tuning	
Scaling	2.0
Stronger Gravity	□
Gravity	1.0
Behavior Alternatives	
Dissuade Hubs	□
LinLog mode	□
Prevent Overlap	□

ForceAtlas 2

Nodes: 31758
Edges: 59291
Undirected Graph

Network Overview		
Average Degree		Run
Avg. Weighted Degree		Run
Network Diameter		Run
Graph Density		Run
Bridging Centrality		Run
HITS		Run
Modularity	0.815	Run
PageRank		Run
Connected Components		Run
Girvan-Newman Clustering		Run
Leiden algorithm		Run
Node Overview		
Avg. Clustering Coefficient		Run
Eigenvector Centrality		Run
Edge Overview		
Avg. Path Length		Run
Dynamic		
# Nodes		Run
# Edges		Run
Degree		Run
Clustering Coefficient		Run

Figure 12. Before applying color by Modularity Class (see Appearance window in upper left).



Modularity Class

🟪	2	(7.16%)
🟩	1	(6.93%)
🟦	4	(4.4%)
⬛	3	(4.02%)
🟧	0	(2.94%)
🟥	6	(2.88%)
🟩	13	(2.44%)
▨	626	(1.69%)
▨	12	(1.51%)
▨	7	(1.35%)
▨	14	(1.32%)
▨	8	(1.17%)

ForceAtlas 2

Threads	
Threads number	7
Performance	
Tolerance (speed)	1.0
Approximate Repulsion	☑
Approximation	1.2
Tuning	
Scaling	2.0
Stronger Gravity	□
Gravity	1.0
Behavior Alternatives	
Dissuade Hubs	□
LinLog mode	□
Prevent Overlap	□

ForceAtlas 2

Nodes: 31758
Edges: 59291
Undirected Graph

Network Overview		
Average Degree		Run
Avg. Weighted Degree		Run
Network Diameter		Run
Graph Density		Run
Bridging Centrality		Run
HITS		Run
Modularity	0.815	Run
PageRank		Run
Connected Components		Run
Girvan-Newman Clustering		Run
Leiden algorithm		Run
Node Overview		
Avg. Clustering Coefficient		Run
Eigenvector Centrality		Run
Edge Overview		
Avg. Path Length		Run
Dynamic		
# Nodes		Run
# Edges		Run
Degree		Run
Clustering Coefficient		Run

Figure 13. After applying color by Modularity Class

The Gephi tool allows color palettes to be automatically generated as well; although this feature is not covered in detail here, the application of the modularity statistic runs along with custom color palettes can be combined to shade the network in whatever number and type of colors will serve the analysis most usefully.

Labelling and Filtering the Graph

Another powerful feature of Gephi that is useful for this analysis is the ability to label nodes and display those labels on the graph. The Data Table has a column called "Label" which is initially blank but Gephi can copy data from one column to another in the Data Table.

Copying the match names into the Label column and turning labels on for the entire network is not particularly useful since it results in an unreadable picture (see Figure 14).

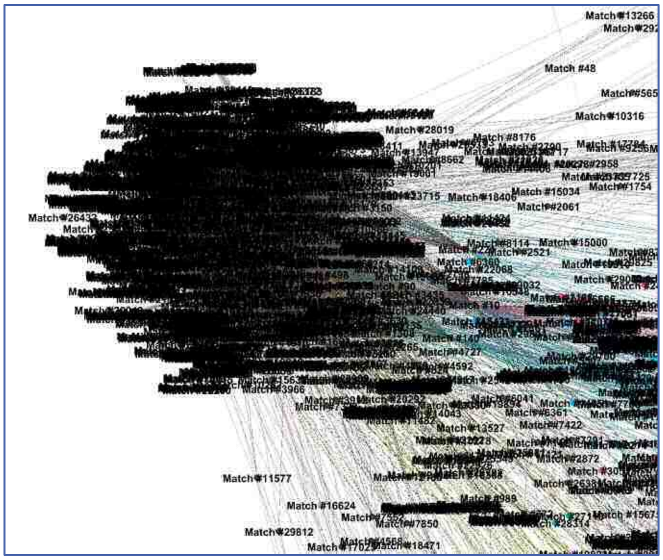
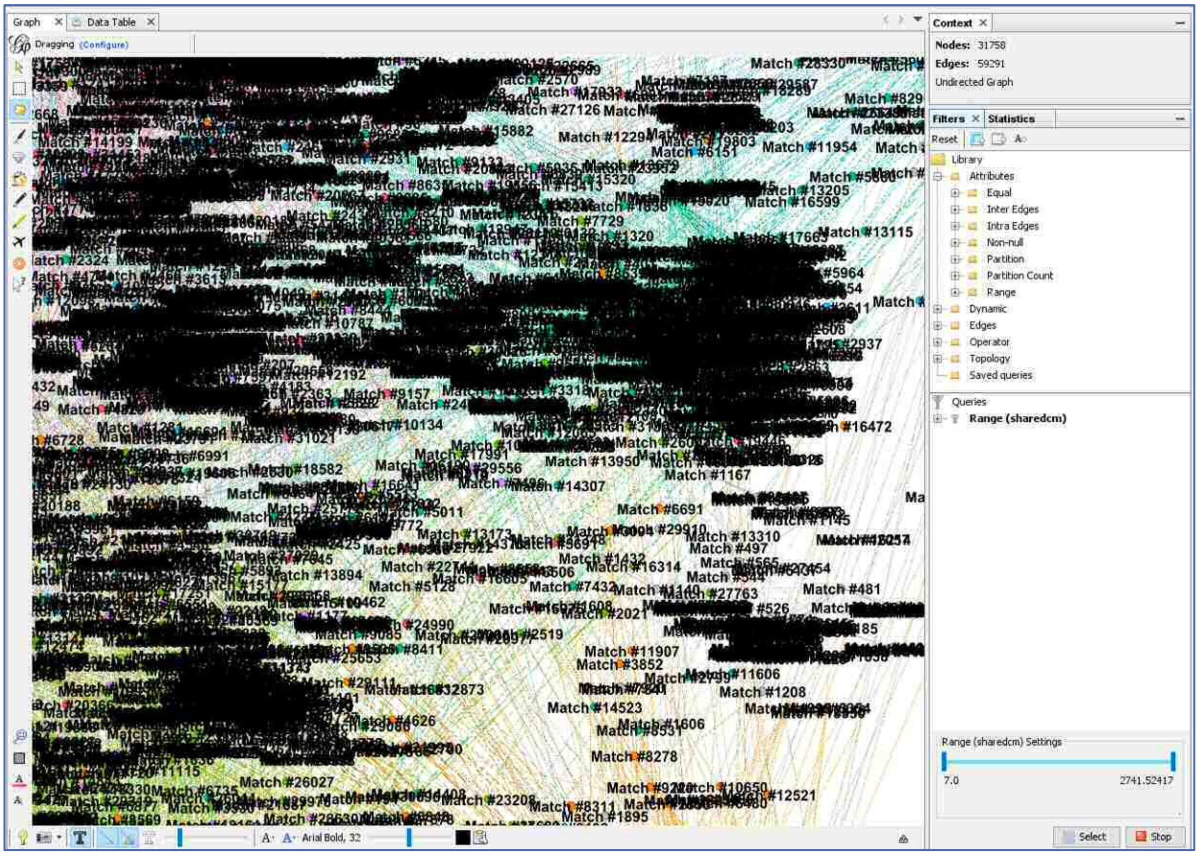


Figure 14. Part of the entire network with labels "on".

However, we can apply filtering to weed out unwanted detail and focus on subsets of the full network.

The Filter options in Gephi can be found on the right-hand side in the companion tab to the Statistics options. From here a variety of filters can be "dragged" down into the Queries window and applied to the network. For example, Figure 15 shows the full network with labels with a Range filter on Shared cMs dragged into the Queries window. This shows that the range of Shared cMs among the nodes ranges from 7.0 at the lower end up to 2741.52417 for the closest match. Changing the lower end of this range to 20.0 and applying this filter will eliminate any nodes (from the graph, not the data table) with less than 20.0 shared cMs . As Figure 16 shows, the display becomes much more readable.



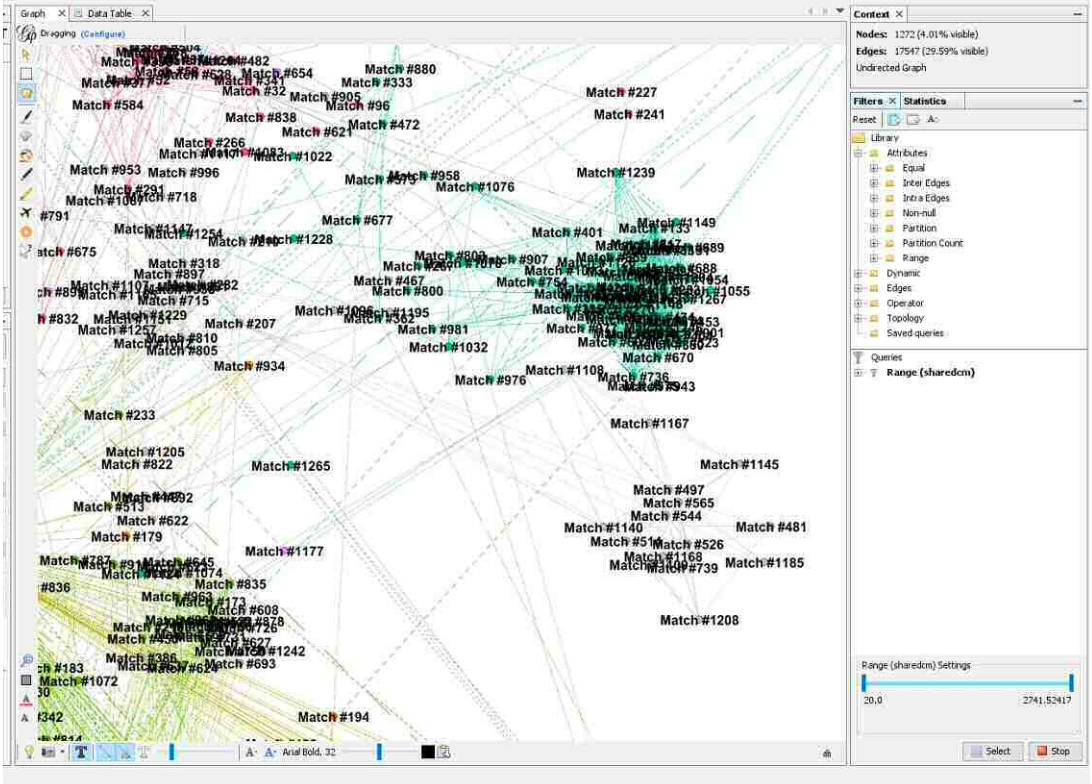
Nodes: 31758
Edges: 59291
Undirected Graph

Library
- Attributes
  - Equal
  - Inter Edges
  - Intra Edges
  - Non-null
  - Partition
  - Partition Count
  - Range
- Dynamic
- Edges
- Operator
- Topology
- Saved queries

Queries
- Range (sharedcm)

Range (sharedcm) Settings
7.0 2741.52417

Figure 15. A Filter on Range of Shared cMs originally selected, showing the existing range among the nodes.



Nodes: 1272 (4.01% visible)
Edges: 17547 (29.59% visible)
Undirected Graph

Range (sharedcm) Settings
20.0 2741.52417

Figure 16. The graph after a filtering has been applied to limit the lower range to 20.0 shared cMs.

Exploring the graph with the nodes labeled by match name may be somewhat useful if clusters can be identified by common surnames or recognizable matches, but mapping other genetic and genealogical information on the network can gain us more insights.

In the Appearance window, we can also vary node size on the display by a data attribute like shared cMs. If we combine this with changing the node labels to the Known Relationships discussed earlier, we get the picture in Figure 17. Note that fewer nodes have labels since we only had Known Relationships for perhaps around 200 of the matches.

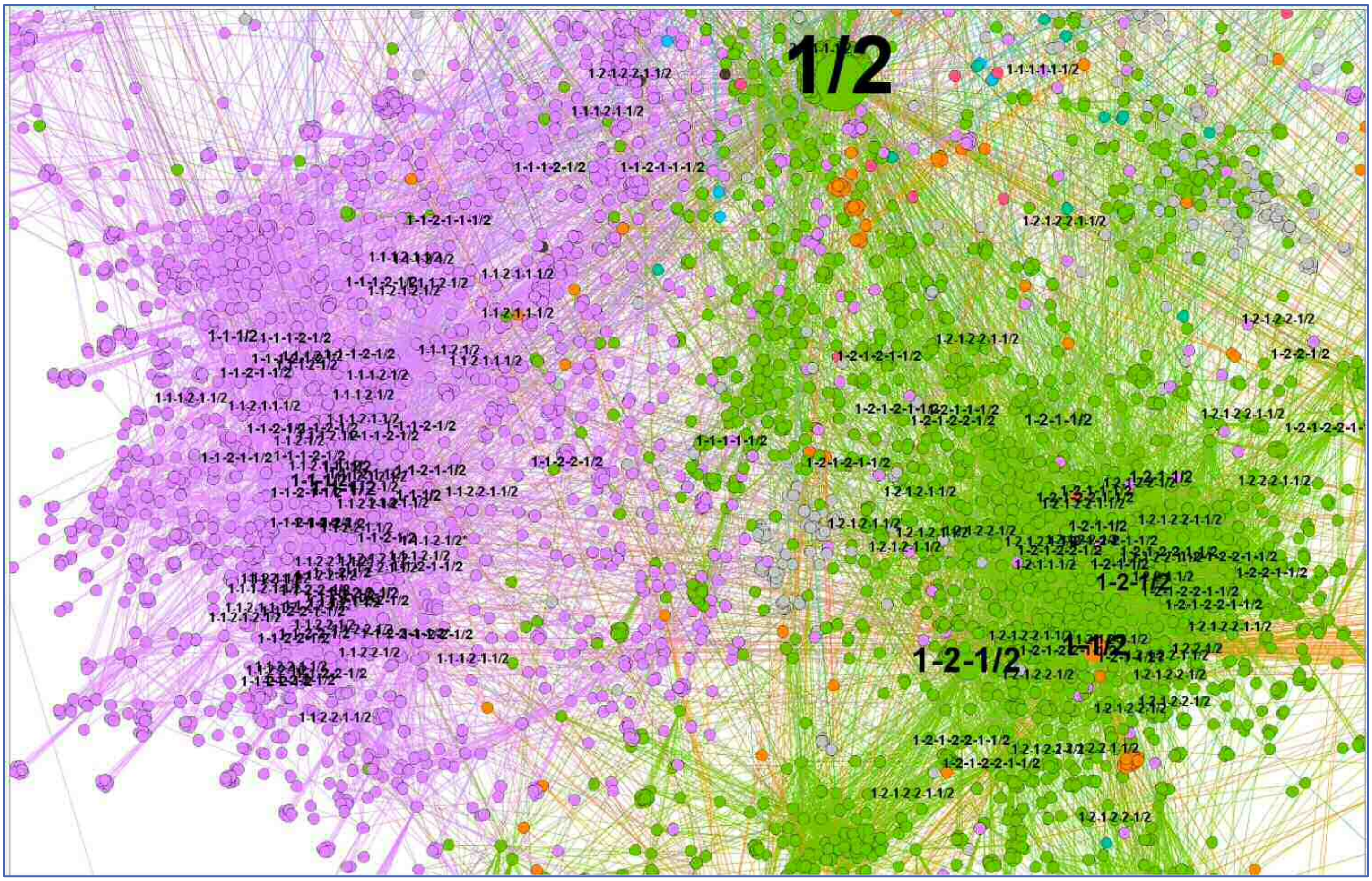


Figure 17. A portion of the network with node and label sizes by shared cMs and Known Relationships used as labels.

Figure 17 is still confusing both from the number of nodes and the influence of the full sibling node at the top of the figure. Removing this sibling node and filtering to eliminate matches in the 7 to 20 cMs range of shared DNA gives us Figure 18:

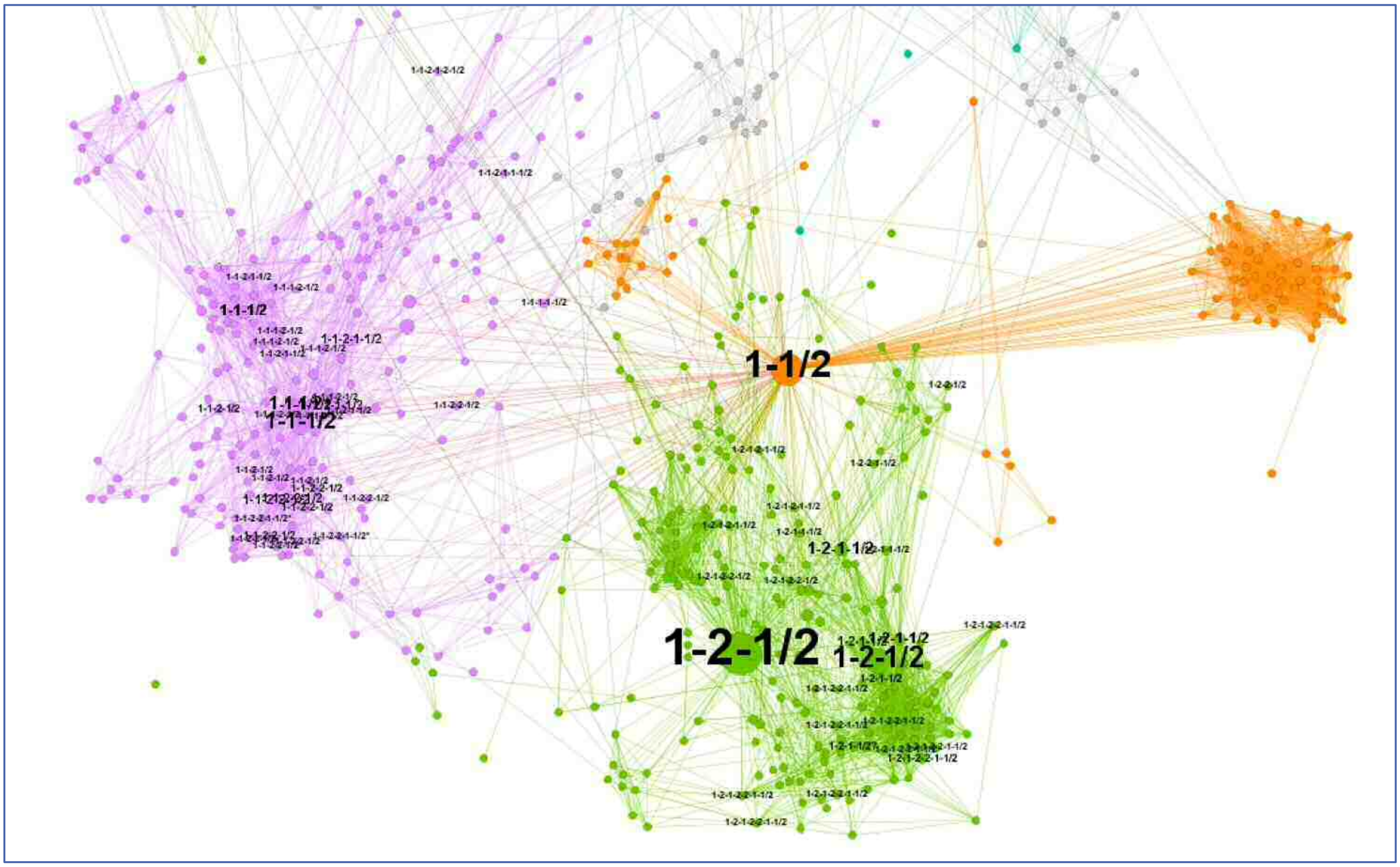


Figure 18. Clusters of nodes above 20 cMs without the full sibling node.

While we could continue to remove the full cousin and full second cousin nodes shown in Figure 18, the origin of these clusters is relatively clear. The purple cluster on the left appears to originate from at least the 1-1-x side of the author's ancestry, meaning our paternal grandfather. There are known relationships on both the 1-1-1-x and 1-1-2-x ancestral lines showing in the purple cluster; these might show better cluster definition if the closer relatives were removed, or perhaps the shared DNA is from both sides. The green cluster however appears to connect in our paternal grandmother's ancestry (i.e. 1-2-x), with a heavy bias to the 1-2-1-2-2 ancestral branch. Again eliminating closer nodes might help break up these clusters for analysis.

The other orange cluster in Figure 18 shows no Known Relationships, but a very tight connection among these ICW matches. Given their consistent connection to the first cousin marked as "1-1/2", they clearly connect on our father's-side ancestry but more cannot yet be determined. After finding the match names associated with these nodes, it is likely that any progress on identifying the common ancestor with any of these matches will narrow down the connections for all of them.

The first cousin's (1-1/2 node) multiple connections to the orange cluster is a pattern that can show with many clusters in the network. Figure 19 shows another cluster with two matches with known relationships to the 2-1-2-1-x ancestral path (and with shared cMs of 51 and 56 cMs), connecting to a cluster formed of nodes with cMs between 20 and 27 cMs.

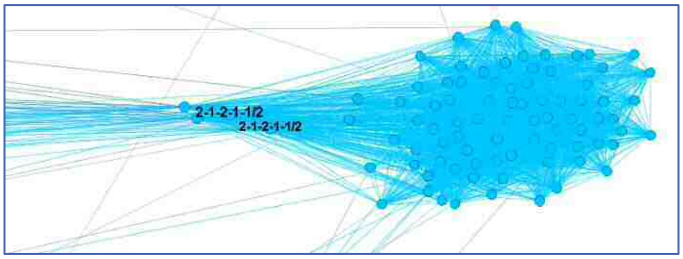


Figure 19. Another cluster in the network

While again endogamy, pedigree collapse and pile-up regions can cloud or even resemble this pattern, it often also is due to closer relations with larger cMs who have inherited DNA from the same older ancestry as the cluster – i.e. that this cluster in Figure 19 has a common ancestor in the generations before our 2-1-2-1 ancestor's parents (in other words, either our mother's father's mother's father's father or mother). Narrowing down these matches at least this far in our common ancestry should significantly reduce the necessary research to identify our specific common ancestors.

In some cases of course, the common ancestry with these matches which connect to a larger cluster will not be known. In those cases these nodes can help focus our research priorities, since identifying the ancestral line of these "connecting nodes" means that the matches in the cluster are also likely connected by ancestors along the same ancestral line.

Making Charts to Share with Family

Separate from analysis (or sometimes to help explain it to others), Gephi can also produce good quality network charts like a simple example in Figure 20 which (if well-explained) can help explain the results of autosomal DNA analysis to other family members as well.

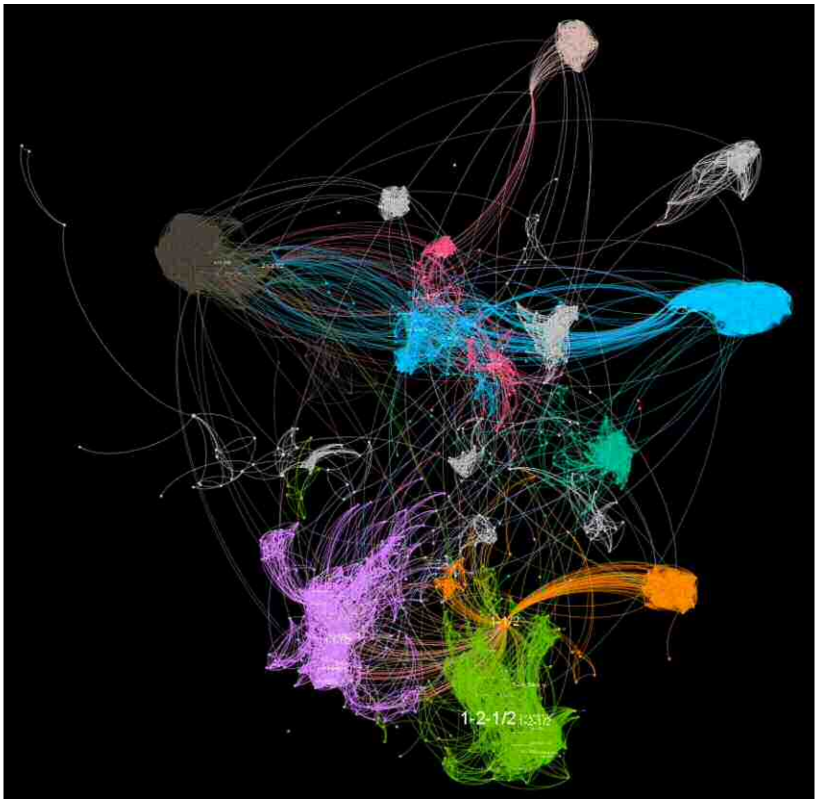


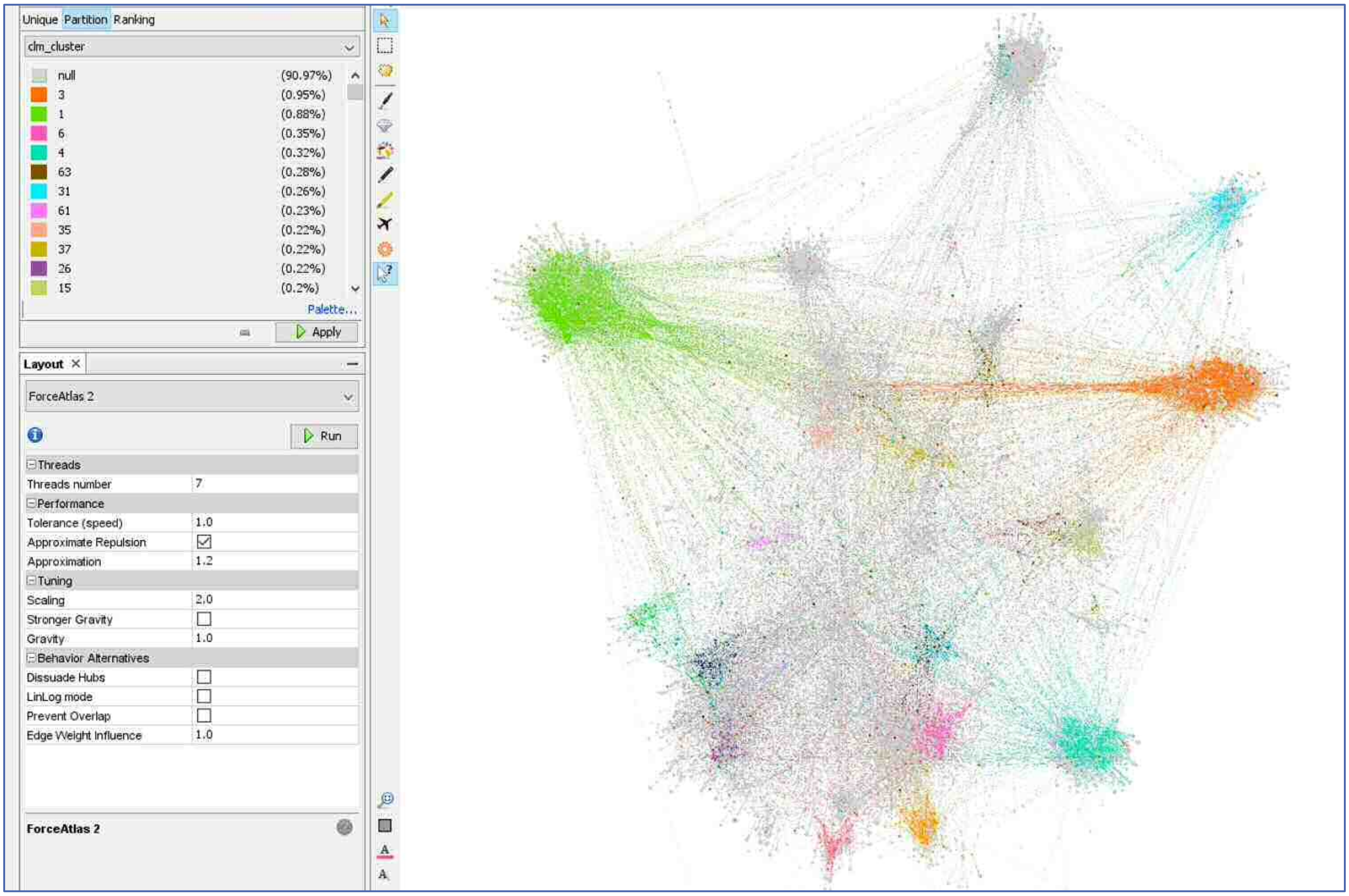
Figure 20. Small changes in color and display can result in shareable views

Comparison to Other Methods

The Collins-Leeds Method (CLM) analysis mentioned earlier was used to generate cluster squares organizing 2,870 of the matches. This was done solely for purposes of this review to provide a comparison of Collins-Leeds Method clustering with Force Atlas 2 clustering.

To perform this comparison, the CLM square numbers for the 2,870 nodes were copied into the Nodes Table data set. Changing the Appearance window to color the nodes based on this column then provides a visual comparison between the clustering approaches in Figure 21. Grey nodes in this figure represent matches who were not sorted into CLM clusters and therefore have no number in the clm_cluster column.

The comparison performed was only by visual inspection since this was performed only to show general alignment and exact correlation was not required. While this is a subjective comparison and not a statistical one, it is clear from Figure 21 that the CLM squares align well with clusters produced by the Gephi analysis.



clm_cluster

⬜	null	(90.97%)
🟧	3	(0.95%)
🟩	1	(0.88%)
🟪	6	(0.35%)
🟩	4	(0.32%)
🟫	63	(0.28%)
🟦	31	(0.26%)
🟪	61	(0.23%)
🟧	35	(0.22%)
🟨	37	(0.22%)
🟪	26	(0.22%)
🟩	15	(0.2%)

ForceAtlas 2

⊟ Threads	
Threads number	7
⊟ Performance	
Tolerance (speed)	1.0
Approximate Repulsion	☑
Approximation	1.2
⊟ Tuning	
Scaling	2.0
Stronger Gravity	□
Gravity	1.0
⊟ Behavior Alternatives	
Dissuade Hubs	□
LinLog mode	□
Prevent Overlap	□
Edge Weight Influence	1.0

ForceAtlas 2

Figure 21. Coloring the CLM Clusters demonstrates a 1-to-1 equivalence with network clusters.

Splitting the Network into Paternal and Maternal Matches

In theoretical models of ancestry, inherited DNA is cleanly split between the portion inherited from a father versus from a mother. While the occurrence of matches will be influenced more by number of children along ancestral lines and the somewhat random chance of lines surviving to present day and being tested, one would still assume that a network of ICW matches would break cleanly into two unconnected super-sets of clusters representing our separate paternal and maternal DNA ancestry. In practice, the split is rarely that clean given the confounding influences mentioned previously. However, by changing the modularity filter's resolution until only two colors remain, we can at least see how easily the network might split into two sub-networks.

To illustrate this, Figure 22 shows the network after the modularity filter has been run with a large enough resolution that only two major colors appear (in this example a resolution of 20.7 was used).

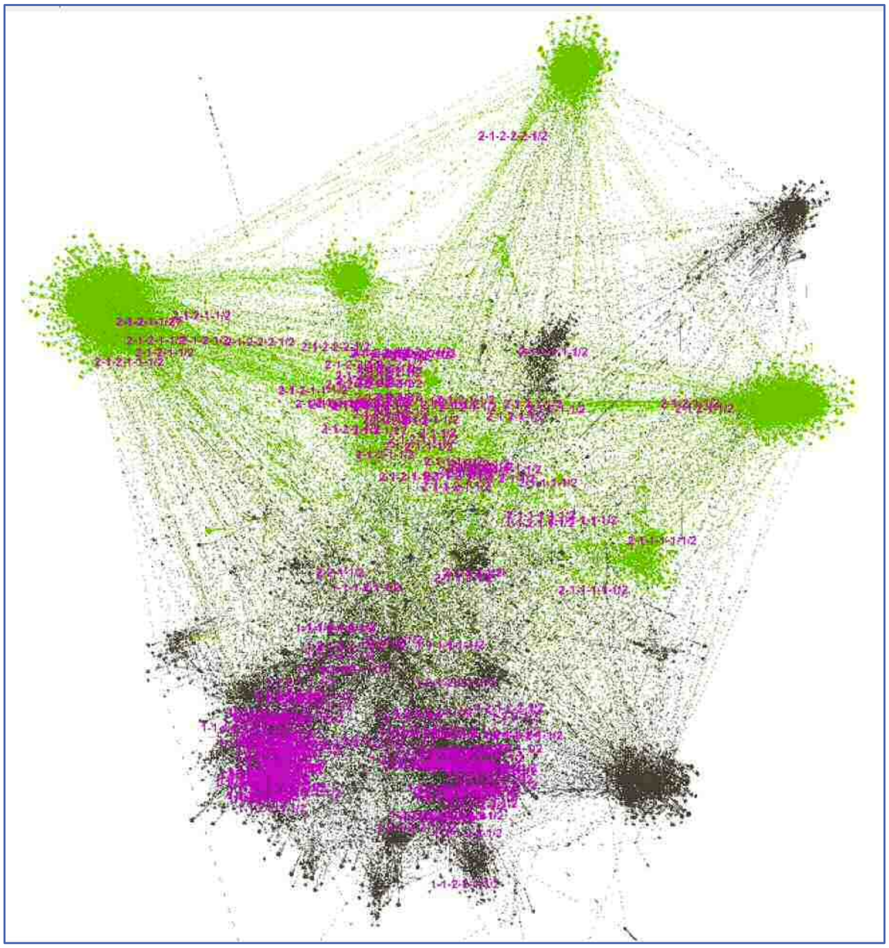


Figure 22. Splitting the Network into two super-sets (in this case, green and black)

For this purpose network visualization is particularly powerful since the identification of these super-sets does not require any knowledge of whether certain matches are paternal or maternal although without any knowledge we could still not assign the super-sets to one side of our ancestry or the other. In Figure 22 however we have increased the size on the Known Relationship labels to look at this more closely. In Figures 23 and 24 the close-ups of Figure 22 show that for our known matches, the green nodes are more clearly from our maternal side and the black nodes from our paternal side.

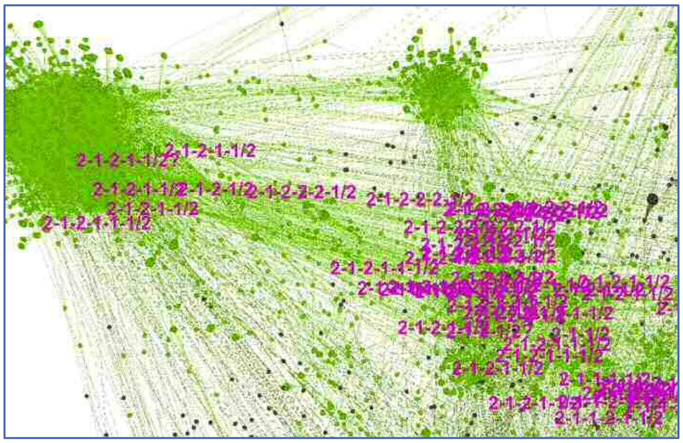


Figure 23. The green nodes are from our maternal side (2-x)

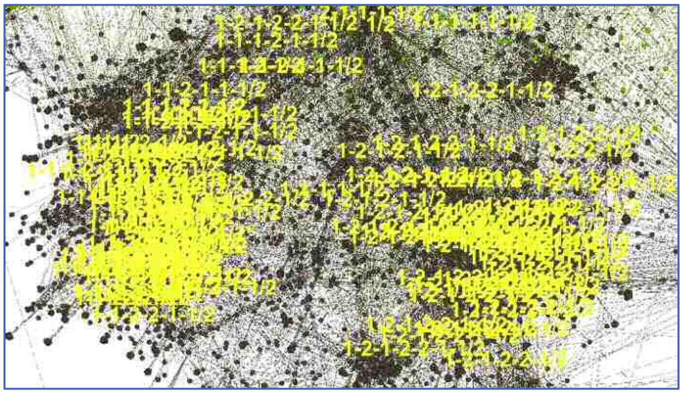


Figure 24. The black nodes are from our paternal side (1-x). (Label colors changed for clarity)

Note that this approach may inadvertently pull some matches from one side over to the other because of connecting edges that confuse the analysis, so especially for nodes that do not aggregate cleanly into large clusters, this split should be taken as suggestive, not conclusive for all nodes.

Looking more closely at the inhibitors to whether a network of ICW matches will split into paternal and maternal relatives, the largest influences may be from the levels of shared endogamy and/or pedigree collapse across the tester's paternal and maternal ancestries since higher levels of either will cause more significant cross-connections that would inhibit the identification of separate sub-networks. DNA pile-up regions may also cause some cross-connectivity between clusters. If the tester's full siblings were also present in the network their nodes would also show significant cross-connections to both paternal and maternal sub-networks; however these nodes can be relatively easily identified and eliminated. Half-siblings and any degree of cousins in the network would not inhibit the identification of paternal and maternal sub-networks since they would (at least based on their primary relationship to the tester) only have connections to one sub-network and not the other.

However even if based only on the potential for endogamy and/or pedigree collapse, not every ICW network will be this easily divisible into paternal and maternal sub-networks. In some cases this may only provide general clues as to whether an unknown cluster is more likely to be from a tester's paternal or maternal sides, but in other cases the exercise may be inconclusive.

If the exercise does identify paternal and maternal sub-networks, one could envision repeating the exercise on each subnetwork to further sub-divide each into grandparent sub-networks and even further. It is however unlikely that any tester's ICW network will subdivide cleanly over several generations, and for that analysis half-siblings or cousins who share ancestry back to those generations would definitely need to be removed from the network.

It is also possible that a similar approach combined with some knowledge of the known origins of specific clusters could be used to identify areas of the network which were affected by endogamy and/or pedigree collapse and so could localize instances of either in the tester's ancestry. This has not yet been studied but remains a potential extension of network visualization analysis.

Extending the Analysis

This review only covers some approaches for using network visualization to focus the likely ancestral lines shared by match clusters to assist in identifying common ancestors. These approaches could easily be extended; for example to include more detailed information from Leeds Method analysis, or details of shared segments by chromosome; or different filtering options could be applied to restrict the network to ranges of shared cMs that correspond to specific limits of generational differences. Once an analyst has some familiarity with the platform used to visualize and filter the network, it can provide a myriad of analysis techniques along with real-time flexibility to switch between them.

Supplementary Information

The author has a video demonstrating these techniques along with more introductory explanations and more detail on working with Gephi options. This video can be found on YouTube at https://youtu.be/Z2T_7aSL4ng.

Conflicts of Interest

The author declares no conflicts of interest and no commercial interest in the topics covered in this review.

References

Gephi tool website, https://gephi.org/. Accessed October 7, 2021.

DNAGedcom tool website, https://www.dnagedcom.com/. Accessed October 7, 2021.

DNAPainter tool website, https://dnapainter.com/. Accessed October 7, 2021.

Genetic Affairs tool website, https://geneticaffairs.com/. Accessed October 7, 2021.

Plos One, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". https://doi.org/10.1371/journal.pone.0098679. Accessed October 7, 2021.

Nicholson, Brit. "Auto-Clusters in Gephi Using Data from GEDmatch". https://dna-sci.com/2020/08/19/auto-clusters-in-gephi-using-data-from-gedmatch/. Accessed October 7, 2021.

YouTube video, "Autosomal ICW Match Analysis using the Gephi tool", https://youtu.be/Z2T_7aSL4ng. Accessed October 7, 2021.

Journal of Genetic Genealogy

Y-DNA SNP-based TMRCA calculations for Surname Project Administrators

by James M Irvine, Administrator, Clan Irwin Surname DNA Project (jamesmirvine@hotmail.co.uk)[1]

Summary

The calculation of a "Time to Most Recent Common Ancestor" (TMRCA) using relevant SNP data can be so simple that many genealogists are tempted to use this tool and to draw inaccurate, imprecise and unwarranted conclusions, however unintentionally. Conversely the calculation of the Confidence Intervals (CIs) that should accompany such calculations is complex and rarely attempted. This paper is not promoting some new panacea, but draws in part on a novel analysis of 17 samples of SNP counts to help genealogists to understand why the popular use of SNP-based TMRCAs without CIs is misguided, why in practice these CIs are difficult to calculate, how curious genealogists can readily estimate indicative CIs for their own data, and why a growing number of genealogists are recognising that the inherent uncertainties which CIs quantify are so great that SNP-based TMRCAs are usually of much less practical use than is often assumed. The development of practical models that include inputs of STR and historical data can reduce these uncertainties, but the temptation to use and mis-interpret simplistic SNP-based TMRCA calculations is not going to disappear.

Introduction

The use of DNA data to calculate TMRCAs is a long-standing objective of the genetic genealogy community. The advent of Next Generation Sequencing (NGS) Y-DNA tests, such as FamilyTreeDNA's BigY test and the resulting SNP haplotrees, appear to offer a significant step towards this goal: many see SNP data as being more reliable than STR data, the calculation of SNP-based TMRCAs can be very simple, published SNP mutation rates are accompanied by a 95% confidence measure which, if not fully understood, seems to add comfort, and above all the resulting TMRCAs appear to be "in the right ball park". Though adopting very different approaches, both Dave Vance's SAPP model[2] and Iain McDonald's recent paper[3] appear to build on these lay perceptions and to supplement the basic SNP-based model with STR and historical inputs. However, the application of such models is too demanding for many administrators ("admins") of surname DNA projects, and as noted in his Conclusion, McDonald has not addressed some of the associated practical problems.

This paper addresses challenges presently facing project admins with limited mathematical skills and/or small samples when they attempt to estimate TMRCAs from SNP inputs derived from NGS tests such as BigY700. The following challenges are addressed in turn:

1. Understanding the basic maths, confidence intervals, accuracy and precision
2. Understanding average SNP mutation rates
3. Understanding SNP counts
4. Practical examples of SNP-based TMRCA calculations
5. Future developments

1. The basic maths

Calculating a predicted TMRCA can be very straightforward, even for those averse to mathematics. What we are exploring here is the use of SNP inputs (only)[4] to predict when the Most Recent Common Ancestor (MRCA) of two or more NGS testers was alive.[5]

[1] I am grateful to Robert Casey, Michael Cooley, Zack Doherty, Maurice Gleeson, David Hall, Jane Lindsay, Tom Little, Kathy McCauley, Dave Vance, Mary Wiley and Dennis Wright who kindly supplied data I used in Appendix B below. I am also very appreciative of comments on earlier drafts of this paper made by Maurice Gleeson, Iain McDonald, Ralph Taylor and Dave Vance, although my adoption of most of their valuable contributions does not imply they necessarily agree with my methodologies, opinions or conclusions.

[2] www.jdvsite.com, accessed 3 November 2021.

[3] McDonald 2021 at https://www.mdpi.com/2073-4425/12/6/862.

[4] I am only addressing SNP inputs to TMRCA calculations in order to keep the paper focussed and relatively short: introducing STR and historical data inputs, while clearly desirable, alas adds further complexity.

Conventionally TMRCAs are based on the years of birth of the tester(s) and of the MRCA. The former is usually assumed to be 1950CE = AD1950.[6] Thus for a single tester:[7]

TMRCA = (year of birth of tester) - (time t in years since the MRCA (a.k.a. coalescence age)

= AD(1950 - t), (1)

where t = r * n years, (2)

where r = relevant mean ("average") SNP mutation rate,
and n = count of SNPs since the MRCA.

For example, if r = 83 years per SNP and the tester has 5 Private SNPs since his Terminal SNP,[8]

t = 83 * 5 = 415 years; and in this case, where the ancestor characterised by this Terminal SNP,

TMRCA = AD1950 - 415 = AD1535.[9]

More generally, if several testers have descended from some MRCA,

t_{mean} = (mean SNP mutation rate) * (mean count of SNPs since MRCA) = $r_{mean} * n_{mean}$ (3)

But while this equation is mathematically correct, it gives a very deceptive impression of both accuracy and precision.[10] It is inaccurate because there are two basic methods of counting the number of SNPs since the MRCA, an issue addressed in section 3.3 below, and it is imprecise because SNP counts vary from patriline to patriline.[11] In fact both the components of equation (3) are Probability Distribution Functions (PDFs) of samples which contain inherent uncertainties. These uncertainties are conventionally represented by three basic equations, one for the estimated average or mean TMRCA,[12] which gives a single "point" date, and two which relate to the precision of this mean value and define its associated uncertainty: one for what is known as the Lower Confidence Interval (LCI) and one for the Upper Confidence Interval (UCI), thus

$t_{LCI} = r_{LCI} * n_{LCI}$ (4)

$t_{UCI} = r_{UCI} * n_{UCI}$[13] (5)

Confidence Intervals, like TMRCAs, are estimates, aka predictions. They are expressed in % terms, such as 95%, i.e. there is a 95% probability that the LCI and UCI calculated from a sample encompass the true TMRCA of the full population, and a 5% probability that they do not.[14]

Another important feature is that the product of two PDFs independent of one another is a PDF which will have CIs of 90% if these CIs are derived from the 95% CIs of the original PDFs,[15] thus:

$t_{90\%} \approx r_{95\%} * n_{95\%}$ (6)

And the product is a PDF with c.50% CIs if these are derived from the 68% CIs of the original PDFs:

$t_{50\%} \approx r_{68\%} * n_{68\%}$[16] (7)

[5] This paper assumes a living person took the NGS test, as opposed to the testing of human remains ("ancient DNA").

[6] CE = Common Era = AD; BCE = Before Common Era = BC. Ethnicity studies usually use "ybp" (years before present).

[7] MRCA is defined by two or more testers, but the TMRCA of a single tester can be calculated if the MRCA SNP is known.

[8] FTDNA effectively defines Private SNPs as those SNPs which have not yet been identified for any other tester, and the Terminal SNP as the most recent SNP currently shared by more than one tester.

[9] YFull assumes testers are aged 60 (www.yfull.com < FAQ). McDonald (2021, 22) cited a sample poll which suggests testers had a mean age of 64 years and a 95% CI of 35-91 years. I gather this poll was taken in about 2014 and so has no impact on mean TMRCA estimates. Unlike the uncertainties in SNP mutation rates and SNP counts which have to be multiplied, these CIs are only additive, and as they are so small in the context of the other uncertainties, in practice they can be ignored. However the "drift" from dates such as these will become increasingly relevant as the years go by.

[10] Accuracy is the closeness of observed data the true value; precision is the closeness of repeated data to each other.

[11] SNP counts also vary considerably within each patriline (aka "lineage"), as shown by McDonald, the Clan Irwin Surname DNA project, and the MacAuley data used in Appendix B below.

[12] For Normal PDFs (only), average = mean = mode = median.

[13] Confusingly t_{LCI} gives TMRCA $_{UCI}$ and t_{UCI} gives $TMRCA_{LCI}$, but in practice this paradox does not affect the results.

[14] A Confidence Interval, aka confidence limit, is twice the margin of error.

[15] Conversely, if the 95% CIs of the product are required, these can be derived using the 97.5% CIs of the original PDFs.

It is important to recognise that equations (6) and (7) are both valid ways to quantify the same uncertainties: they simply express them in different forms. But which form is the more appropriate for TMRCA calculations in the genealogical context? The use of 95% CIs is customary in mathematics and some sciences,[17] and has been adopted by some genealogists. But as shown above, 95% CIs for SNP rates and for SNP counts give 90% CIs for TMRCAs, and there are several reasons why in practice 95% CIs for SNP counts are less predictable than has been appreciated hitherto, and why 68% CIs for SNP counts and the consequential 50% CIs for TMRCAs are more appropriate than 95/90% CIs:

- Appendix A below shows that few SNP counts have a clear "best fit" PDF, and the 68% CIs of the Normal, Poisson and Log-Normal PDFs differ from one another less than their respective 95% CIs, thus making the choice of PDF gives the "best fit" PDF as less critical;
- It also shows that unless the sample has a large number of testers, the 95% CIs derived from the actual cumulative frequencies of SNP counts are less reliable than similarly-derived 68% CIs;
- TMRCA 50% CIs span narrower date ranges than 90% CIs, and so are less likely to be irrelevant to conventional genealogical research (although of course they are also irrelevant 50% of the time);
- Fewer project admins should dismiss all CIs as academic niceties and/or remote contingencies.[18]

Another mathematical issue is the precision of the results of TMRCA calculations, and specifically the extent to which their results can or should be rounded. Given the value of TMRCA calculations currently perceived by many genealogists this is an important matter, although there seems to be no consensus. It is possible to calculate TMRCAs and their associated CIs to the nearest year[19] but as we will see, given the extensive number of uncertainties, several of which are not yet quantifiable, it seems appropriate to round off the results of calculations of TMRCAs to, say, the nearest 10 or even 50 years, and the associated lower and upper CIs to even more, perhaps to the nearest 100 years.[20]

While it is desirable to "round off" TMRCAs to avoid unjustified precision, this does not mean that the two components of SNP-based TMRCAs (SNP mutation rates and SNP counts) should only be calculated and cited to two significant figures, as this practice introduces avoidable and confusing errors, albeit that such errors may be trivial in the context of the underlying uncertainties.

2 Average SNP mutation rates

First, some more simple maths. Genetics theory tells us that

Average SNP mutation rate, r, in years per SNP = 1/(base pairs frequency * base pairs length) (8)

Several data sets are relevant:

Data set	base pairs' frequency	base pairs' length (hg38)	r, years per SNP
YFull (ComBED coverage)	$8.2*10^{-10}$ bps/year	8,482,579 bps	144.41
BigY500 (ex Warehouse, accessed 3 Nov. 2021)	($8.2*10^{-10}$ bps/year)	9,286,211 bps	131.32
FGC Elite 1 (ex Warehouse, accessed 3 Nov. 2021)	$8.2*10^{-10}$ bps/year	14,007,575 bps	87.06
BigY700 (ex Warehouse, accessed 3 Nov. 2021)	($8.2*10^{-10}$ bps/year)	14,626,759 bps	83.34
McDonald approximations (2021,3,23)	$8*10^{-10}$ bps/year	15,000,000 bps	"83" (83.33)

This table makes clear the important feature that there is no single "correct" average SNP mutation rate r: McDonald (2021, 3, 23-24) explains that the appropriate average rate depends on the length, measured in base pairs ("bps"), of the male-specific portion of the Y chromosome and on the

[16] The product of two 68.3% CIs is a 47% CI, but for practical purposes I am assuming it to be 50% - an "evens" likelihood that the true mean is within these CIs and also, in this example, of it being outside these CIs. For Normal PDFs, 68.3% CIs = mean +/- SD, and 95.4% CIs = mean +/- 2*SD, and so the probability of a 95.4% CI is half that of a 68.3% CI.

[17] For example 95% CIs are appropriate in ethnicity studies, because typical SNP counts and sample sizes are much larger.

[18] These last two reasons are subjective: some may argue that the 50% CI ranges are still too wide to be of practical value to genealogists, and/or that a 50% chance that the CIs are irrelevant makes them valueless.

[19] Until recently Alex Williamson's https://www.ytree.net was expressing TMRCAs to the nearest decimal of a year!

[20] A parallel issue is how CIs are best expressed. Thus a TMRCA can be described as "AD1600, with 90% CIs of +200 years and -300 years" or "between AD1300 and AD1800 (90% CIs), with a mean of AD1600". Some even argue that it is preferable to omit the central year and only give a range of dates with their percentage CIs.

number of SNPs that are "callable". Thus for example if following a BigY test the relevant SNPs being counted are as analysed by YFull then the average rate of 144.4 years/SNP is appropriate, immaterial of the original test,[21] whereas if the SNPs being counted are those listed by FTDNA then the YDNA-Warehouse rates of 131.3 years/SNP are appropriate for BigY500 test results, and 83.3 years/SNP for BigY700 test results.[22] McDonald's rate of 83 years/SNP is an approximation, the absence of any decimal places no doubt being deliberately intended to reflect the underlying uncertainties.[23]

McDonald (2021, 24) lists a selection of published studies of modern average SNP mutation rates:

Paper	Reference	base pairs' frequencies				95% CIs		68% CIs	
	in McDonald,	mean	95% LCI	95% UCI	skewed?	as ratio		as ratio	
	2021	all in terms of 10^{-10} p.a.				of mean		of mean	
Mendex et al.	[27]	6.17	4.39	7.07	left	0.288	0.146		
Adamov et al.	[5]	7.98	6.32	9.84	slightly right	0.208	0.233		
Poznik et al.	[30]	8.2	7.2	9.2	symmetric	0.122	0.122	0.06	0.06
Helgason et al.	[21]	8.33	7.57	9.17	slightly right	0.091	0.101		
Xue et al.	[28]	10	3	25	left	0.700	1.500		

From this he concludes that the callable SNP mutations in the Y-chromosome have a probable mean of c.$8.2*10^{-10}$ base pairs p.a., and that the distribution about this mean is probably a Poisson PDF.

In section 2.2.2 of his paper McDonald considers different methods for handling the uncertainties of the SNP mutation rates for multiple tests. For his objective of a general model which also includes inputs of STR and historical data he treats the uncertainties in the mutation rates as he computes each SNP node in turn. For simplicity, and with less sophistication, this paper instead first explores the implications of recognising these uncertainties as a scaling factor over each cohort of SNPs. In this latter context the above data can be extrapolated thus:[24]

Analysis company/ databank	Test	Basis for Confidence Intervals	mutation rate r, years per SNP				
			Mean	95% CIs		68% CIs	
				Lower	Upper	Lower	Upper
YFull	various	Adamov	144.4	120.6	178.1	128.5	160.3
FTDNA / Warehouse	BigY500	Poznik	131.3	115.3	147.3	123.4	139.3
FGC / Warehouse	Elite 1	Poznik	87.1	76.4	97.7	81.5	92.3
FTDNA / Warehouse	BigY700	Poznik	83.3	73.2	93.6	78.4	88.4

If the TMRCA being sought is to be based on SNP data for patrilineal descendants who have taken a variety of NGS tests, for example some BigY500 and some BigY700, then the SNP count for the Private SNPs in the BigY500 or BigY700 sample will need to have a correction factor applied.[25]

3 SNP counts

Four issues have to be considered when developing the relevant count of SNPs which are to be multiplied by the average mutation rates developed above:

1. Recognition that SNPs vary in quality and should be counted consistent with the test coverage.

[21] Note that YFull analyse FGC, FTDNA BigY500 and FTDNA BigY700 test data, but for all data they only count SNPs in the comBED regions before using their 144 years/SNP rate to calculate TMRCAs; in contrast the relevant rates published by YDNA Warehouse are applicable whenever the counts of SNPs called by FGC or FTDNA are used to calculate TMRCAs.

[22] See https://ydna-warehouse.org/coverage.html, accessed 3 November 2021. Note that the Warehouse data is updated from time to time and so the most up-to-date rates may differ slightly. For example, when accessed on 22 July 2021, BigY700 showed r = 83.38 years per SNP.

[23] For an alternative approach see www.jdvsite.com/faq < link to video on analysing BigY matches using SAPP.

[24] I have generated the mutation rates in this table by extrapolating the CIs in the previous table, which for the Poznik data curiously implies a symmetric PDF, not a Poisson PDF. McDonald used the Helgason data, which has slightly narrower CIs. Note also that the mutation rate relevant to an individual test depends on both the type of test and the exact coverage of the individual test, and I understand the latter can vary by about 10%. I am assuming, perhaps naively, that the 95% CIs in all these papers address the aggregate of all the relevant the uncertainties.

[25] Alternatively the data may be 'normalised' to some uniform bps length (e.g. as with the Royal Stewart data in Appendix B). Or the less reliable sample may be ignored (e.g. most would argue Y500 data is less reliable than Y700 data).

2. Recognition that the SNP counts currently available from NGS testing for each patriline are only a sample of those of the wider population of testers descending from their common ancestor;
3. Recognition of the bias inherent in most mean counts of SNPs since the MRCA
4. Recognition of the variety of methods of calculating the associated Confidence Intervals.

3.1 Recognition of the varying quality of Individual SNPs. While SNPs are not prone to the convergency/back mutation issues that bedevil STR data, and so can be considered much more stable, nevertheless what does or does not constitute a novel, callable, "phylogenetically significant" SNP is not clear-cut and unambiguous for every SNP. The quality of the callable SNPs to be counted depends on several factors, including:

- Coverage. The SNPs being counted must all occur in the same region of the Y chromosome as that where the chosen SNP mutation rate has been validated.
- Mapping. SNPs can be mis-called by the sequencer if two repeats are misaligned with each other
- Depth. For a SNP to be considered callable it must have a minimum number of reads, aka calls, typically 4, overlapping the location, but there is no fixed standard.
- Read consistency. Typically at least 90% of the reads should be derived rather than ancestral.

These features are beyond the scope of this paper, but the crucial point is that the criteria used for the SNP count should be the same as that used for the relevant test: if the criteria are different this may have a significant impact on TMRCA calculations.

FTDNA, YFull, Alex Williamson in his BigTree[26] and McDonald in his 2021 paper all consider the relevant VCF data or even BAM data for each SNP and make their own judgements as to what does or does not constitute a callable SNP.[27] There is an assumption that the FTDNA data stored in the YDNA Warehouse databank represents a consistent judgement on which SNPs are callable.

The project admin (or individual tester) has three options for handling this issue. They can:

- send the VCF or BAM data for the relevant test(s) to YFull (or to some private "expert"), bearing in mind that (i) although the Y-Full fee ($49) for a single NGS test is relatively trivial, the cost quickly increases as more project members subscribe, and (ii) the utility of this option is dependent on how many other matching NGS test results are already in relevant sections of the YFull haplotree; or
- analyse the VCF or BAM data themselves, as for example done by McDonald, or as explained by Vance.[28] This is a most satisfying and illuminating exercise, but difficult and laborious for the untrained, and involves project admins having to justify why their own analyses differ from those of YFull and/or FTDNA; or
- follow FTDNA's determination of which SNPs are callable for each individual BigY test in light of their ever-increasing awareness of the haplotree of mankind.[29] FTDNA frequently refine these details, which means that the haplotree and Private SNPs that appear on their web pages are kept updated as the haplotree matures, but are liable to change from time to time.

The vast majority of testers and project admins choose the third of these options, albeit perhaps unconsciously.

But whichever option is chosen, there is a need to periodically review the SNP count that has been used to calculate a TMRCA.

3.2 Recognition that SNP counts are samples of a larger population. While it is tempting to regard SNP counts as deterministic, and to assume the TMRCA back to some designated SNP that can be calculated without any probability caveats such as CIs, most TMRCA calculations involve more than one tester sharing descent from some MRCA SNP, and the SNP count frequencies are inherently some form of PDF, even if it is not necessarily a close fit with one of the more common continuous

[26] See https://www.ytree.net. Dennis Wright has similarly processed the O'Brian and R-L226 data in Appendix B below.

[27] The ISOGG Y haplotree has yet another set of criteria for what constitutes a phylogenetically-significant SNP.

[28] See www.jdvsite.com/faq < link to video on analysing BigY matches using SAPP.

[29] This is additional to their routine "naming" of previously un-named SNPs that were "Private" to another tester.

PDFs. Nor are such SNP counts static, for they evolve over time as more such descendants take the relevant NGS test, and as the "callability" of marginal SNPs evolves in response to improved understandings of SNP quality, as described above). In fact the available relevant SNP counts of most patrilines are samples of larger populations of descendants, typically of unknown size, and this introduces inherent uncertainties in the available evidence of SNP counts.

3.3 Recognition of the bias inherent in most mean counts of SNPs. The use of the mean count of SNPs since the MRCA when calculating TMRCAs, as in equation (3) above, is a convenient and popular method.[30] In contrast YFull and McDonald[31] apply a more refined method of counting SNPs since the MRCA, using an unbiased node-by-node SNP count. This method involves a sequential, bottom-up counting process for each node of the haplotree in turn. The methodology is best understood by way of an example – see section 4.1 below. If the relevant haplotree is fully symmetrical in shape then the two methods will give the same resultant count of SNPs since the MRCA, but of course all but the very simplest of haplotrees are asymmetrical and so in practice the two methods give different SNP counts.[32] Several issues thus arise:

- Why do the two methods give different results? This is because the mean SNP count method introduces a bias at each node in the haplotree if the patrilines below this node represent different numbers of testers; for example, if there are three patrilines below a node, of which one represents three testers, one represents two testers and the other represents a single tester, then by the mean count method the first two patrilines give undue weight to the SNP counts of their respective testers. Again this is best understood by the example in section 4.1 below. It follows that this node-by-node method avoids the biases that are inherent in averaging such data, and is thus clearly more accurate than the more convenient mean SNP count method.
- Why then is use of this unbiased node-by-node method not more popular? The biases that this method avoids have only been publicised relatively recently and so few project admins are aware of it, and of those who are, some do not recognise its greater accuracy. It is also more complicated and laborious to apply, and, especially for large sample sizes or if the haplotree is not redrawn to show the nodes clearly, is more prone to errors during application.[33] It can also be argued that the difference in the results of the two methods is likely to be well within the associated Confidence Intervals,[34] and so the extra effort is not justified. Nor, because of its relative complexity, is this node-by-node likely to become more popular than the mean count method in the future, unless its complexity can be circumvented by some user-friendly software.
- How do the results of the two methods differ? Because the difference for each sample will depend on the shape of each relevant haplotree it is not possible to predict the size of the difference, or even to develop some simple "rule of thumb" to forecast which method will give the larger SNP count, as the following examples show:[35]

Sample			Sample size	SNP count		Difference		
			(no. of testers)	convenient mean	node-by-node	SNPs	%	TMRCA
Royal Stewart	S781	var	26	5.96	6.86	-0.90	-13%	-113 years
Border Irwin	FGC13746	Y700	65	7.57	6.72	+0.85	+13%	+71 years
Lae/Lay	FT21692	Y700	34	3.56	2.79	+0.77	+28%	+64 years
Doherty	BY472	Y500	30	10.38	10.83	-0.45	- 4%	-60 years
MacAuley	Y179697	Y700	32	7.59	6.88	+0.71	+10%	+59 years
Doherty	BY471	Y700	63	12.90	13.60	-0.70	- 5%	-58 years
Irwin	FT104360	Y700	6	13.33	12.83	-0.50	- 4%	-42 years

[30] For examples of this method see Holton GD *Tracing your Ancestors using DNA*, Barnsley 2019, pp.137-9, and Dave Vance's SAPP model; the latter uses a hybrid method: although SNPs are counted at each node, the mathematical result is the mean SNP count, to which is applied an arbitrary weighting to address outliers (see www.jdvsite.com/faq < link to video analysing BigY matches using SAPP, accessed 3 November 2021). Note both sources caution that the results of their TMRCA calculations should include a margin of error of a couple of centuries either way.

[31] www.yfull.org; McDonald 2021.

[32] The two methods also give the same result if the TMRCA is between a single tester and one of his ancestors.

[33] This will remain true until the process can be encapsulated into a sophisticated computer program.

[34] This statement is correct, although the differences relate to accuracy whereas the CIs relate to precision.

[35] This data was collected before I had appreciated the significance of biases attributable to the haplotree shape.

More study is needed to clarify this issue, but meanwhile we can note that these differences imply that mean count TMRCA calculations may incorporate errors of +/- c.1-3 generations.

3.4 Recognition of the variety of methods of calculating associated Confidence Intervals. Although it has long been recognised that calculations of predicted TMRCAs should always be accompanied by their associated CIs, and that such CIs can be characterised, at least conceptually, by some common type of PDF, in practice such CIs are very rarely calculated. There are several reasons for this: few individual testers and project admins are interested; some argue that they know the range of CIs is so large that it would be a waste of time to calculate them; a small minority who are curious find the methodology unclear and the maths too complicated; and even the most diligent analysts struggle to access samples of SNP counts that are usually not publicly available. And, I now appreciate, it was naïve to expect that SNP counts can be reliably represented by one of the common PDF types. So hitherto little attempt has been made to use empirical data to determine how SNP-based TMRCA CIs should be derived. This subject is developed further in Appendix A below, whose findings in this context may be summarised thus:

- There is no single, "text book" method with which SNP-based TMRCA CIs should be calculated.
- Even large SNP counts cannot be represented by any common PDF because of "noise" attributable to features such as asymmetrical haplotrees, population size, family size, father's age etc.
- Of the more common PDFs, the Poisson PDF is the "best fit" (though not necessarily a "good fit") to the samples of SNP counts within the surname era[36] that have been analysed.
- The CIs associated with any "best fit" PDF are not necessarily the most reliable guide to the CIs associated with such counts. CIs can also be interpolated from the actual cumulative frequencies of the SNP counts in the sample, except when the number of testers is small, these CIs can offer a more reliable method than the CIs derived from "best fit" PDFs.
- So conceptually the choice of the appropriate CIs for each specific sample of SNP counts could be determined by identifying the PDF giving the best "best fit" to this data, then comparing the CIs derived from this "best fit" PDF with the CIs derived from the actual cumulative frequencies of the SNP counts, and finally making a subjective choice of the most appropriate CIs.

Although pragmatic, such a conceptual process has many disadvantages: it is laborious and error-prone, it is impractical for TMRCA calculations when fewer than c.15 testers share a MRCA SNP, it is unattractive to genealogists lacking the necessary patience or understanding of statistical theory, and after all the effort the chosen CIs are not necessarily as objective or accurate as the calculations imply. To avoid all but the last of these issues I have developed two simple, empirical formulae which can be readily used by non-mathematically minded genealogists for all SNP-based TMRCA calculations within the surname era, even when SNP counts of only a few testers are available:

Estimated Lower CI$_{68\%}$ of SNP count ≈ mean SNP count - (sq.root mean SNP count * TF$_L$) (8)

Estimated Upper CI$_{68\%}$ of SNP count ≈ mean SNP count + (sq.root mean SNP count * TF$_U$) (9)

where TF$_L$ and TF$_U$ are factors derived from the following table for 68.3% CIs:[37]

No. of testers (N):	2	3	4	5	6	7	8	9	10	12	15	20	30	50	100	∞
TF$_{L\ 68.3\%}$:	1.85	1.32	1.20	1.14	1.11	1.09	1.08	1.07	1.06	1.05	1.04	1.03	1.02	1.01	1.01	1.00
TF$_{U\ 68.3\%}$:	2.775	1.98	1.80	1.71	1.67	1.64	1.62	1.61	1.59	1.58	1.56	1.55	1.53	1.52	1.52	1.50

These formulae are used to estimate the CIs on the bottom line of the entries for each sample in column O of the Summary table in Appendix B. The resulting CI estimates can be seen to be pretty close to the CIs derived from the "best fit" CIs and those interpolated from the cumulative frequencies; the few that are narrower are shown in *italic* font. This implies that however crude and

[36] By surname era I mean the period since when surnames first became hereditary, typically c.600-1,000 years ago.

[37] In this table TF$_L$ is copied from the similar table introduced in Appendix A below. TF$_U$ is simply (TF$_L$ * 1.5), where 1.5 is an arbitrary, empirical factor to allow for the longer right-tail of the Poisson PDFs which Appendix A has shown to be characteristic of most SNP counts within the surname era. Similar simple formulae could readily be developed for 95% CIs, but the simplifications would introduce appreciable errors and also give a misleading impression of accuracy.

illogical these simple formulae may be, they nevertheless offer a fair "rule-of-thumb" indication of the uncertainties associated with SNP counts, at least until a more refined substitute is developed.

So subject to various assumptions already addressed, equations (3) - (6), (8) and (9) above can be modified to give three simple, ubiquitous equations, thus:

$$\text{TMRCA}_{\text{mean}} = \text{AD}(1950 - (r * n)) \tag{10}$$

$$\text{TMRCA}_{\text{50\%LCI}} \approx \text{AD}(1950 - (r_{\text{68\%UCI}} * (n + (\text{sq.root } n * \text{TF}_{\text{U}})) \tag{11}$$

$$\text{TMRCA}_{\text{50\%UCI}} \approx \text{AD}(1950 - (r_{\text{68\%LCI}} * (n - (\text{sq.root } n * \text{TF}_{\text{L}})) \tag{12}$$

Clearly equations (11) and (12) only offer approximate estimations and are no substitute for mathematical rigour.[38] The three equations can be represented by the following "look-up" table:[39]

Table for estimating TMRCAs based on mean SNP counts and associated indicative 50% Confidence Intervals for BigY700 testers

Year AD of TMRCA & 50% CIs		No. of testers (N) descended from MRCA											
		2		4		6		10		30		∞	
Mean SNP count since MRCA (n)	Predicted TMRCA	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI
1.0	1867	1616	2017	1702	1966	1714	1959	1721	1955	1726	1952	1729	1950
2.0	1783	1426	1998	1548	1926	1565	1916	1574	1911	1582	1906	1586	1904
3.0	1700	1260	1966	1409	1878	1430	1866	1441	1859	1451	1853	1455	1851
4.0	1617	1106	1926	1278	1825	1302	1810	1315	1803	1326	1796	1331	1793
5.0	1534	959	1882	1152	1768	1179	1753	1194	1744	1206	1737	1211	1733
6.0	1450	819	1835	1030	1710	1059	1693	1075	1683	1088	1675	1095	1672
7.0	1367	682	1785	910	1650	942	1631	959	1621	973	1613	980	1609
8.0	1284	549	1733	793	1589	826	1569	845	1558	860	1549	868	1545
9.0	1200	418	1680	677	1527	713	1505	733	1494	749	1484	757	1480
10.0	1117	290	1625	563	1464	601	1441	622	1429	638	1419	647	1414
11.0	1034	164	1569	450	1400	489	1376	511	1363	529	1353	538	1348
12.0	950	39	1512	338	1335	379	1311	402	1297	421	1286	430	1281
13.0	867	BC84	1454	227	1270	270	1245	294	1230	313	1219	323	1213

NB Above dates are given to nearest year to avoid interpolation errors;
after interpolating for intermediate values of mean SNP count and no. of testers,
dates of TMRCAs should be rounded to nearest 10 years, and
dates of 50% Confidence Intervals should be rounded to nearest 100 years

Assumptions: Testers born AD1950; mean SNP mutation rate: 83.3 yrs per SNP, 68% CIs 78.4 - 88.4 yrs/SNP;
Lower 68% CI of mean SNP count n = sq.root n*(t factor); Upper 68% CI = 1.5*Lower CI.

This table enables curious genealogists to derive indicative CIs associated with TMRCAs derived from mean SNP counts, and to keep such TMRCAs in perspective. But more importantly it demonstrates clearly that even at best, 50% of the time the mean TMRCA will be outside a CI range of at least two centuries, and typically of much longer periods. And of course 90% CIs cover even wider periods.[40] This confirms the unpalatable and widely unrecognised fact that, pending possible future developments (see section 5 below), TMRCAs derived from SNPs alone are of much more limited practical use to genealogists than many of them assume.

Strictly speaking equations (11) and (12) and the above table are not applicable if TMRCAs are calculated with the node-by-node method to avoid biases, although it could be argued that the difference between mean-based and unbiased node-by-node based SNP counts is a matter of accuracy whereas the associated CIs are a matter of precision. On the other hand the calculation of CIs for node-by-node based TMRCAs would be tortuous without a software package.

[38] For example, in theory CIs scale as sq.root n/(N-1), but this is offset by the SNP counts not being fully independent.

[39] The shaded dates are of Upper 50% CIs which are later than the mean birth dates of the testers. Similar tables could be developed for BigY500. Predicted TMRCA accuracy is improved by the node-by-node method (see section 3.3 above).

[40] The same can be said for TMRCAs based on STRs alone.

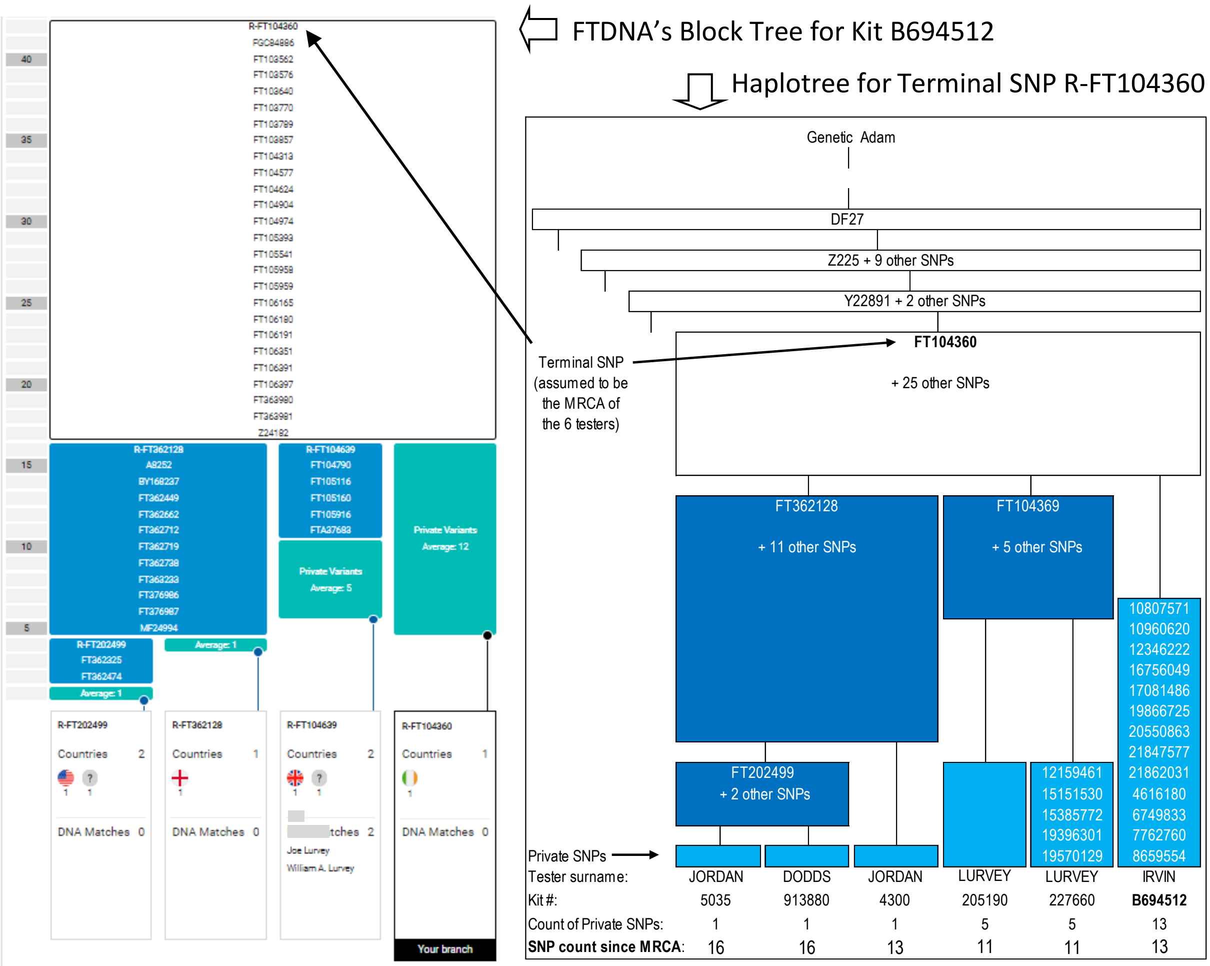
4 Practical examples of SNP-based TMRCA calculations

The following examples use data from the Clan Irwin Surname DNA project[41] to illustrate some practical applications of the above considerations when calculating TMRCAs from SNP data alone.

4.1 Predicting TMRCAs from data in FTDNA's Block Trees

This is a common challenge for FTDNA's project admins. Below is a copy of the Block Tree for a Clan Irwin tester kit B694512 who has only one YDNA "match" in the Irwin Surname DNA project, who in turn happens to be a tester with the surname Lurvey. The question arises of whether one of these two "matches" has an NPE[42] in his ancestry, or whether their shared Terminal SNP, identified by FTDNA as R-FT104360, is so old that it probably represents a common ancestor who lived before the surname era.

The first step is to aggregate as much relevant SNP data as possible[43] from publicly available sources to build the haplotree downstream of the Terminal SNP R-FT104360.[44] By referring to FTDNA's public web pages for the Jordan, Dodds and Lurvey surname projects, the Block Tree for kit B694512 (below, left) can be expanded into a haplotree for SNP R-FT104360 (below, right):


FTDNA's Block Tree for Kit B694512

Haplotree for Terminal SNP R-FT104360

Tester surname:	JORDAN	DODDS	JORDAN	LURVEY	LURVEY	IRVIN
Kit #:	5035	913880	4300	205190	227660	**B694512**
Count of Private SNPs:	1	1	1	5	5	13
SNP count since MRCA:	16	16	13	11	11	13

All six of these men have inherited all 26 SNPs in the R-FT104360 block, although pending more BigY test results we do not know which of the SNPs in this block is the most recent. Nevertheless if we count the SNPs in each patriline subsequent to this block we can calculate the likely TMRCA of the 6 men thus:

[41] See www.clanirwin-dna.org.

[42] For a discussion of NPEs see www.isogg/org.wiki < Non-paternity event.

[43] The more SNPs that can be included in the TMRCA calculation the less unreliable the calculation will be.

[44] FTDNA's Block trees are limited to 30 matches, and so project admins may have to refer to more than one Block tree to build the relevant haplotree.

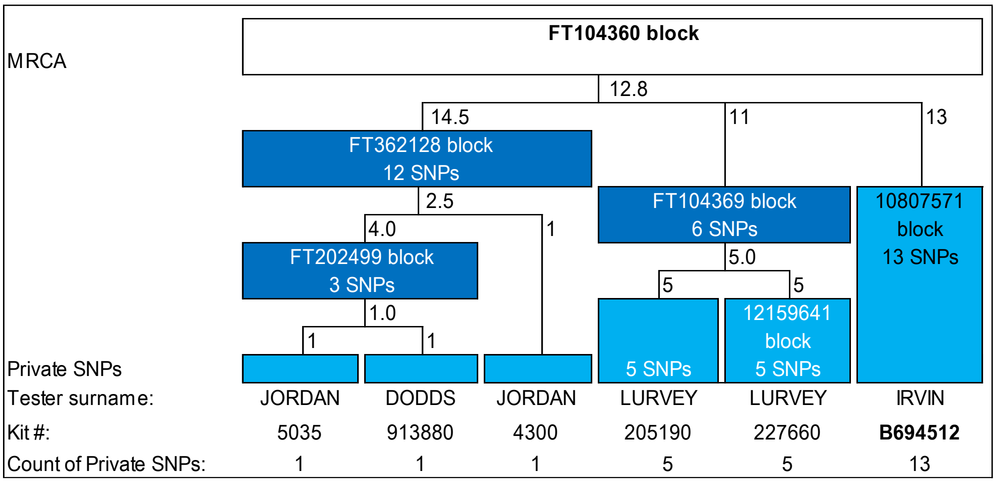
total SNP count = $\sum n = (16 + 16 + 13 + 11 + 11 + 13) = 80$ SNPs[45]

mean SNP count = $(\sum n)/n = 80/6 = 13.3$ SNPs

age to coalescence = $t = r * n_{mean} = 83.3 * 13.3 = 1111$,[46] say 1110 years

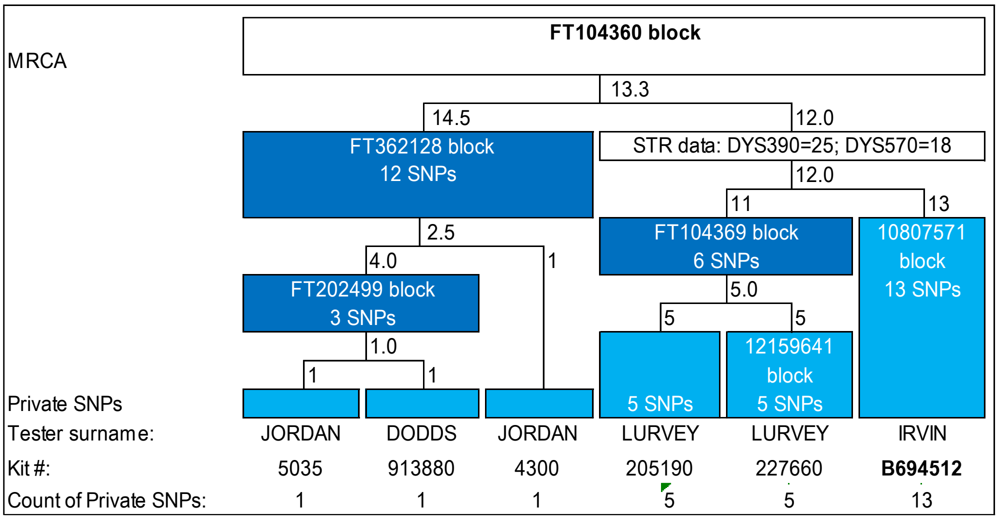
and hence TMRCA ≈ AD1950 - 1110 = AD840.

However this calculation is biased by the dominance of the 3 men sharing descent from FT362128. Such biases are avoided by instead adopting the more refined node-by-node method addressed in section 3.3 above. This method can be applied to this example by adapting the haplotree thus:



The pros and cons of node-by-node calculations versus the more convenient averaging of SNP counts have been developed in section 3.3 above. The implications for this particular example are that the MRCA SNP count changes from 13.3 to 12.8 SNPs, and thus the TMRCA from AD840 to AD880.

A further refinement, albeit outside the nominal scope of this paper, is to adapt this haplotree into a mutation history tree[47] by adding some STR data. Inspection of the FTDNA public pages for these 6 men shows that the Lurveys and Irvin share different counts for two STRs, DYS390 and DYS570, from the other 3 men. This may be represented thus:



The MRCA count thus changes again, by coincidence from 12.8 back to 13.3, and the resulting TMRCAs from AD880 back to AD840. But whether the predicted TMRCA is AD840 or AD880, both these dates predate the surname era, and so it is unlikely that either the Irwin or Lurvey patrilines

[45] Why the "Block Tree" for B694512 shows 12 Private SNPs but his "Results" show 13 is unclear. I note similar minor discrepancies from time to time, and regard FTDNA's "Results" data as being more reliable than their "Block Tree" data.

[46] We know that all six testers took the BigY700 test because they share a terminal SNP prefixed by "FT", which is effectively specific to SNPs identified by BigY700 tests.

[47] A mutation history tree is a haplotree extended to include relevant STR data.

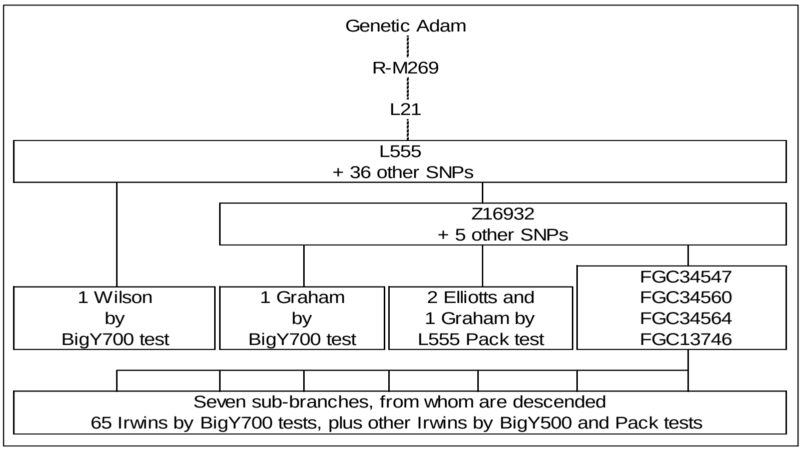
included a NPE that adopted the other's surname.[48] Instead it is more likely that each patriline descending from the R-FT104360 block included one or more SNP mutations which predated the surname era.

4.2 Can we identify the SNP characterising the MRCA within a SNP "block".

The small branch of Irwins descended from R-FT104360 is one of over 40 branches of this surname identified by the project. The largest of these branches is the Borders branch, which currently has 65 BigY700 testers.[49] Some descendants of this branch still live today near the Scottish Borders, and a few have patrilines dating back to ancestors who lived in Dumfriesshire c.1500. Alas we know neither the name nor dates of the founder of this branch, but by combining some historical evidence with the R-L555 haplotree we can identify the SNP "block" which is most likely to include the SNP representing the earliest Border Irwin. The immediate challenge is to see if we can tentatively identify which SNP within that block is most likely to be representative of the MRCA.

Let us first consider the historical evidence. The surname de Irwyn occurs across central Scotland in the 13th century, but the lengthy Ragman Roll of 1296 (in which no Irwins appeared) suggests that the use of hereditary surnames across Scotland was then still only common amongst the nobility. The earliest extant use of de Irwyn as a hereditary surname was in Aberdeenshire during the 14th century. When extant records become prolific in the Scottish Borders in c.1500 we find evidence of several contemporary branches of the surname in Dumfriesshire which appear to be loosely related to each other, suggestive of a common ancestor a few generations earlier. There is some more specific evidence: a Nicholas de Irwin who was briefly a vicar at Buittle in the 1370s, and John and Gilchrist de Irwin who were tenants at Buittle and Morton respectively in 1376.[50] It is possible that Nicholas was the father of John and Gilchrist, and if so that Nicholas was born c.1320, his sons were born c.1350, and it would have been his grandsons born c.1380 who were the first Irwins to be born in Dumfriesshire. But this is speculation, for we have no evidence that Nicholas, John or Gilchrist had surviving offspring, or that one or more were an ancestor of later Border Irwins. The challenge is thus to see if ySNP data can throw any light on these uncertainties.

When the SNP R-L555 was discovered seven years ago by a Walk-the-Y DNA test it seemed possible that this SNP might be unique to the Border Irwins. It is now apparent that this SNP is shared by some Wilsons, Grahams and Elliotts, which like Irwin are surnames common on the western Border, and that it is a block of four SNPs younger than R-L555 which are unique to the Border Irwins:[51]



Pending further NGS tests which may break down this block of four SNPs, we do not know the sequence in which their respective mutations occurred, and so the sequence in which these SNPs

[48] This is (just) so even after considering the Upper 50% CI, in this case c.AD1300 (from table in section 3.4 above).

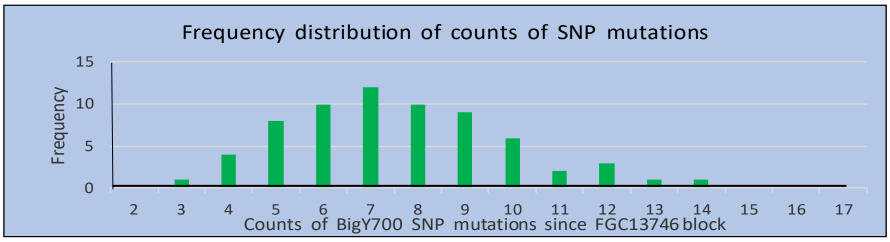
[49] 9 further Border Irwins have taken BigY500, but their inclusion in this section would add unnecessary complication.

[50] Buittle and Morton are 10-15 miles west of Dumfries. Although a Gilchrist filio Eruni was a witness in Dumfries in 1124x1185 and two Irwins lived in Berwick on Tweed c.1330, the former cannot have been a hereditary name, and there is no evidence that the latter had descendants. For more details see Irvine *The Irwin Surname* 2020, 65-9, 139.

[51] "Border Irwins" include various spellings of the name sharing descent from R-FGC13746, and NPEs sharing this SNP.

are listed, whether by FTDNA or anyone else, is quite arbitrary, as is the SNP chosen to identify the block. Nevertheless the TMRCA of descendants of this block can be readily calculated, as can the date that this block was formed.

The following bar chart shows the current count of BigY700 SNPs downstream of R-FGC13746 block:

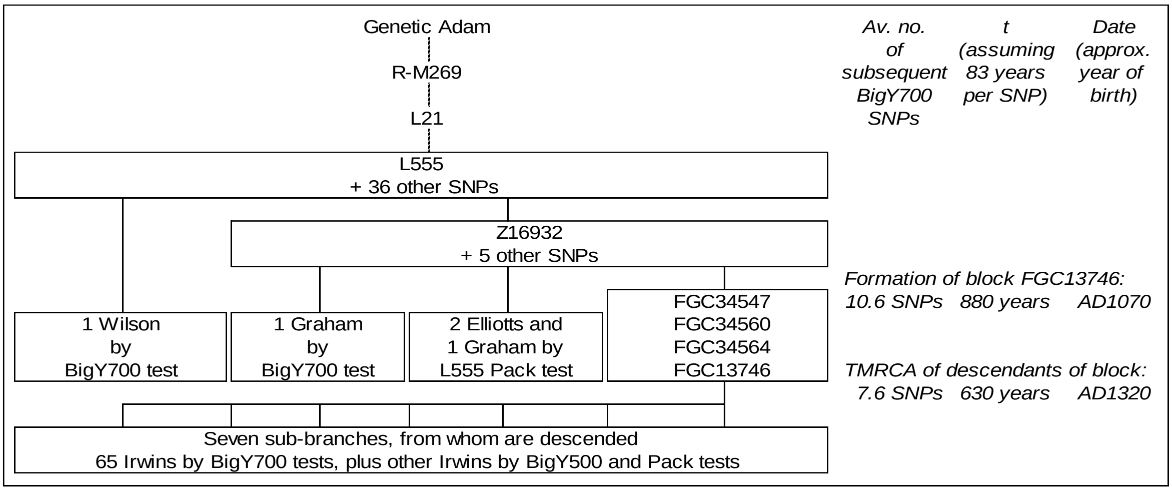


The mean count of SNPs since the FGC13746 block is 7.6,[52] and so

$t_{mean} = n * \mu = 83.3 * 7.6 = 633$ years, say 630, and hence

TMRCA ≈ AD1950 - 630 = AD1320.

Similarly the mean count of SNPs since the earliest SNP in this block is 10.6 SNPs, giving date of the formation of this block as AD1070.[53] We thus have, prima facie:


Genetic Adam
R-M269
L21
L555
+ 36 other SNPs
Z16932
+ 5 other SNPs
1 Wilson
by
BigY700 test
1 Graham
by
BigY700 test
2 Elliotts and
1 Graham by
L555 Pack test
FGC34547
FGC34560
FGC34564
FGC13746
Seven sub-branches, from whom are descended
65 Irwins by BigY700 tests, plus other Irwins by BigY500 and Pack tests
Av. no. of subsequent BigY700 SNPs
t (assuming 83 years per SNP)
Date (approx. year of birth)
Formation of block FGC13746:
10.6 SNPs 880 years AD1070
TMRCA of descendants of block:
7.6 SNPs 630 years AD1320

So if the date range of the FGC13746 block is AD1070 to AD1320, it seems the youngest SNP in this block might represent the earliest Border Irwin, even if we don't know its identity. However the more accurate node-by-node method gives the range of SNP counts for the FGC13746 block as 6.7 to 9.7, and hence the date range as AD1140 to AD1390, suggesting it might be the second youngest SNP in this block that best represents the earliest Border Irwin, even if its identity is also unknown.

It is thus apparent that "No", we cannot determine exactly which SNP within this block characterises the earliest Border Irwin. And while it may seem comforting that the above TMRCAs are at least compatible with the historical evidence,[54] it is important that even this tenuous impression needs to be placed in the context of the associated confidence intervals.

[52] Calculated in Appendix C.

[53] YFull apply the term "formed" to the describe age of the oldest SNP in a block, in this example 880 ybp, or AD1070.

[54] It is curious that this Border Irwin example of a SNP-based TMRCA matching closely to historical data is not unique:

Sample		TMRCA based on			
		Historical data only	convenient mean SNP count	node-by-node SNP count	McDonald 2021, 19
Royal Stewart	S781	c.1245	c.1200	c.1090	c.1254
Border Irwin	FGC13746	c.1350	c.1320	c.1390	-
MacAuley	FT22697	c.1766	c.1760	c.1820	-

No doubt research into other surname branches can reveal similar examples. But we have to recognise that such examples are fortuitous coincidences rather than conclusive evidence. It is also coincidental, and ironic, that in these three examples the convenient mean-based TMRCAs appear more accurate than the node-by-node based TMRCAs!

4.3 Estimating approximate Confidence Intervals ("CIs")

Section 3.4 above has shown these calculations can be made quite simple. For example, using the Border Irwin R-FGC13746 data introduced above and equations (10), (11) and (12) above we have:

TMRCA$_{mean}$ = AD(1950 - (r$_{mean}$ * n)), where r$_{mean}$ = 83.3 yrs/SNP and n = 7.57 SNPs
= AD(1950 - (83.3 * 7.57)
= AD(1950 - 631)
≈ AD1320[55]

TMRCA$_{50\%LCI}$ ≈ AD(1950 - (r$_{68\%UCI}$ * (n + (sq.root n * TF$_U$)), where r$_{68\%LCU}$ = 88.4 yrs/SNP & TF$_U$ = 1.52
≈ AD(1950 - (88.4 * (7.57 + (2.75 * 1.52))
≈ AD(1950 - 1034)
≈ AD920

TMRCA$_{50\%UCI}$ ≈ AD(1950 - (r$_{68\%LCI}$ * (n - (sq.root n * TF$_L$)), where r$_{68\%LCI}$ = 78.4 yrs/SNP (ex section 2),
sq.root n = 2.75, and N = 65, so TF$_L$ = 1.01 (ex section 3.4)
≈ AD(1950 - (78.4 * (7.57 - (2.75 * 1.01))
≈ AD(1950 - 376)
≈ AD1570

i.e. TMRCA = AD1320 (50% CI: 920-1570).

Similar dates can be interpolated from the table at the end of section 3.4 above.

For those who are averse to using approximate tools such as these, Appendix C shows that the "best fit" PDF for this data is a Poisson PDF, and entering 7.57 as the number of "Observed Events" at www.statology.org/poisson-confidence-interval-calculator gives 68.3% CIs of 4.9 and 11.5, so we have:

TMRCA$_{50\%LCI}$ = AD(1950 - (r$_{68\%LCI}$ * n$_{68\%LCI}$)
= AD(1950 - (78.4 * 4.9)
= AD(1950 - 384)
≈ AD1570

and TMRCA$_{50\%UCI}$ = AD(1950 - (r$_{68\%UCI}$ * n$_{68\% UCI}$)
= AD(1950 - (88.4 * 11.5)
= AD(1950 - 1016)
≈ AD930

i.e. TMRCA = AD1320 (50% CI: 930-1570)

and Appendix C also shows that the actual cumulative frequencies give 68% CIs of 4.6 and 10.4, so:

TMRCA$_{50\%LCI}$ = AD(1950 - (r$_{68\%LCI}$ * n$_{68\%LCI}$)
= AD(1950 - (78.4 * 4.6)
= AD(1950 - 361)
≈ AD1590

and TMRCA$_{50\%UCI}$ = AD(1950 - (r$_{68\%UCI}$ * n$_{68\% UCI}$)
= AD(1950 - (88.4 * 10.4)
= AD(1950 - 919)
≈ AD1030

i.e. TMRCA = AD1320 (50% CI: 1030-1590).

This latter method for calculating TMRCA CIs is probably the most reliable for this sample, and for most samples with mean SNP counts of > c.15, but the first method is the easiest to calculate, and is also the only method applicable to samples with mean SNP counts of < c.15. But even with all this attention we almost certainly have not yet captured all the subtleties of SNP mutations, so it is probably safest just to say that for this sample:

TMRCA = AD1390 (50% CI: c.900-1600).

While CIs such as these "flag" the uncertainties inherent in all SNP-based TMRCA models, this example exemplifies the warning in section 3.4 above that even 50% CIs render TMRCAs based on SNPs alone to be of very limited practical use to the genealogist. It is thus understandable that some project admins advise that all SNP-based TMRCA calculations should be taken with a large "pinch of salt", or even entirely ignored.

[55] As we have seen, the more accurate node-by-node method gives a TMRCA of AD1390 for this data.

5 Future developments

This paper is not suggesting that this disappointing demonstration of the severe limitations of SNP-based TMRCAs is the last word on this subject, or that the empirical equations (11) and (12) for estimating TMRCA CIs are a substitute for mathematical rigour. Several future developments can be identified that should help to narrow the CI ranges that measure the uncertainties inherent in all TMRCA calculations:

- Both Vance and McDonald have already demonstrated the ability to combine SNP inputs with STR and historical data inputs for calculating TMRCAs for genealogists. Both these models represent substantial advances, even if at present the former does not avoid the biases inherent in asymmetrically shaped haplotrees and lacks quantified CIs, while the latter is still accompanied by practical problems, and its inherent complexity makes the associated maths too advanced for use by most genealogists. Significant refinements in both models are thus clearly feasible, albeit the challenges remain formidable.
- FTDNA are expected to soon improve their on-line resource for deriving TMRCAs and thus hopefully contribute a significant step forward for most project admins.
- The pending YDNA Warehouse website platform will hopefully enable much more extensive and rigorous analyses of empirical SNP data than has been possible in Appendix A to this paper, thus leading to improvement in the understanding of the characteristics of various features that may influence SNP mutations.
- New DNA testing technologies and analysis tools may also help refine TMRCA calculations.

Notwithstanding these likely developments, SNP data will continue to contribute to TMRCA calculations, the need for accompanying CIs will not cease, and many of the principles addressed in this paper will remain relevant. And nor is the simplistic attraction of multiplying SNP counts by some mean SNP mutation rate to derive misleading TMRCAs likely to disappear.

6 Conclusions

This paper addresses several important conclusions associated with SNP-derived TMRCAs:

1. The mathematically simple equation of TMRCA = AD1950 - (r * n), where r is an appropriate average SNP mutation rate and n is some relevant SNP count, gives a very deceptive impression of both accuracy and precision, unless accompanied by appropriate confidence intervals (CIs).

2. The appropriate r to use, such as 144, 131 or 83 years per SNP, depends on the circumstances.

3. There are two basic methods of counting SNPs when calculating TMRCAs; these may be described as the mean SNP count method and the unbiased node-by-node method. The former is better known and more convenient to use; the latter is more accurate as it avoids biases due to haplotree asymmetry. TMRCAs calculated by the two methods may differ by 1-2 generations.

4. TMRCAs are liable to change over time as more descendants test and as callable SNPs are refined.

5. SNP rates and counts are samples of larger populations and are thus probability distribution functions (PDFs). The more testers whose SNPs can be counted, the more precise their calculated TMRCAs become.

6. TMRCAs calculated using SNP rates and SNP counts with 95% CIs have 90% CIs, and those calculated using 68% CIs have c.50% CIs.

7. For several reasons it is arguable that TMRCAs with 50% CIs (derived from SNP rates and SNP counts with 68% CIs) give genealogists a more appropriate guide than TMRCAs with 90% CIs.

8. TMRCAs should be rounded to the nearest 10 or 50 years and their associated CIs to 100 years.

9. There is no clear "correct" method for calculating CIs of SNP counts. Most SNP counts within the surname-era have a Poisson PDF "best fit", but in many cases this "best fit" PDF gives a narrower range of CIs than those derived from the cumulative frequencies of actual SNP counts, when these can be calculated, i.e. when the number of testers exceeds about 15.

10. Equations (11) & (12) and the table in section 3.4 above offer tools for deriving indicative CIs that are simple for curious genealogists to apply, less prone to calculation errors, and, uniquely, are applicable when the number of testers is small. The validity of these tools is confirmed by their application in Appendix B below.

11. Notwithstanding the above caveats and refinements, the range of CI values associated with SNP-based TMRCA models is often so great that alone these models are of much less practical use than many genealogists believe.

12. Appendix A below also shows that the expectation that the probability distributions of the larger samples would tend towards a smooth bar chart and a close fit to one of the common PDFs that might be expected from the random mutations of SNPs was naive. Instead the “spikiness” of these SNP counts is indicative of “noise” caused by other factors such as biases inherent in the asymmetric shape of most haplotrees, changes in population size, varying family sizes, the age of the father, Founder effect etc.

13. Several potential developments may help alleviate these otherwise depressing findings. These developments could include further refinement of Vance’s SAPP tool, a more practical version of the model recently developed by McDonald, and the pending improvements in the current FTDNA and YDNA Warehouse websites. How much such developments may improve the accuracy and precision of TMRCA calculations remains to be seen, but they will not remove all the underlying uncertainties and nor will they override the underlying principles addressed in this paper, or make the superficial attraction of simplistically derived SNP-based TMRCAs disappear.

14. Pending such developments, all TMRCAs, not least those based on SNP inputs alone, even when accompanied by CIs, remain a very crude tool for the typical genealogist, unless confirmed by a reliable independent source.

Appendix A - Statistical analysis of SNP Counts

To date most genetic genealogists have given little consideration to the various statistical characteristics of the SNP counts that many surname project admins are using to calculate TMRCAs. Such characteristics include the relevance of the number of testers descended from some shared SNP characterising the their Most Recent Common Ancestor (MRCA),[56] the frequencies and mean values of the relevant SNP counts, identifying which (if any) Probability Distribution Function (PDF) gives a "good fit" with these counts, and deriving associated Confidence Intervals (CIs).

This Appendix addresses findings from 17 samples of SNP counts, each with >15 testers sharing descent from a MRCA SNP, which are thought to constitute the first such survey of this subject.[57]

It is necessary to start by recognising that counts of SNPs of men sharing patrilineal descent from a MRCA SNP are but a sample, typically growing over time as more descendants test, of a larger, otherwise untested, population of a particular surname branch or haplogroup. As such the sample is subject to the mathematics of sample probabilities. It is also necessary to recognise that few genealogists are comfortable with the jargon, theoretical aspects and practical complexities of this discipline, even if many still expect that TMRCAs can somehow contribute to their own studies.

In this context it is convenient to first address the relevance of the number of samples I have surveyed and of sample size, i.e. the number of testers within each sample. Common sense tells us that where the number of samples or of testers is small then the statistical uncertainties will be dominant, and conversely that the larger the number of samples or testers available the more reliable our findings can be. In the context of YDNA SNPs, the number of samples in the public domain is small because the SNP counts of testers who share some MRCA SNP requires awareness of what FTDNA term as Private Variants, and this information is only readily available to the individual tester and his project administrator.[58] This is one reason why this subject has remained unexplored hitherto, and why in order to get most of the samples in this survey I had to resort to "citizen science" in my appeal of 8 September 2021 to the Facebook Group "Project Administrators Only". The resulting survey of 17 samples, though small, is illuminating.

Similarly the number of testers within any given sample is often small because Direct-To-Customer (DTC) Next Generation Sequencing (NGS) testing, such as FTDNA's BigY tests, is still relatively new and expensive, and few such testers can yet be grouped as sharing a MRCA SNP within the surname era.[59] Nevertheless determining the impact of the sample size, i.e. the number of testers, on the Confidence Intervals we need for TMRCA calculations is relatively straightforward, and can be illustrated by the following adaptation of Student's "t" factors:[60]

No. of testers (N)	2	3	4	5	6	7	8	9	10	12	15	20	30	50	100	∞
1 sided 84% 2-sided 68.3%	1.85	1.32	1.2	1.14	1.11	1.09	1.08	1.07	1.06	1.05	1.04	1.03	1.02	1.01	1.01	1.00
1 sided 98% 2-sided 95.4%	13.8	4.50	3.29	2.86	2.64	2.51	2.42	2.36	2.31	2.25	2.19	2.14	2.08	2.05	2.02	2.00

This table shows how the contribution of the CIs of the mean SNP count to TMRCA calculations can be modified by a simple factor to take account of the number of testers.

[56] This may be a single SNP, or the youngest SNP within a "block" of SNPs even if the sequence of SNPs within the block, albeit the identity of this youngest SNP is not yet known.

[57] Readers should be aware that I did not devise this survey to imitate, verify or criticise the TMRCA model developed in McDonald's 2021 paper, and I conducted the survey before I had appreciated the inherent biases in data of this type due to features such as the asymmetry of haplotrees, population size etc.

[58] The data in Alex Williamson's excellent https://www.ytree.net does not readily show the type of test (e.g. BigY500 vs. BigY700) for each tester sharing some MRCA, or the number of callable Private SNPs, or, more critically, do his haplotree branches all terminate at the same SNP level. Nor is the even more comprehensive YDNA Warehouse data available in a suitable format, though this may change with its pending move to a new platform.

[59] This Appendix is only addressing SNP counts of DTC NGS tests, and not of ancient DNA samples, or the application of SNP counts to ethnicity or haplogroup studies.

[60] Conventional "t" tables are entered with the number of "degrees of freedom", which equal N-1 where N is the sample size; hence here the number of testers must be 2 or more.

In theory these “t factors” should only be applied to Normal and Log-Normal PDFs and not to asymmetric PDFs such as Poisson, but even if the precise quantifications are not applicable to such PDFs, the underlying principle clearly applies,[61] particularly in the context of 68% CIs.[62] These considerations also explain why, in practice, as the number of testers in a sample slowly grows, the relevance of sample size decreases, and we can focus on other, more fundamental, parameters.

Before analysing the implications of these 17 samples of SNP counts it is first appropriate to develop some hypotheses. After sample size the next most significant parameter is apparently the mean of the SNP counts in each sample. If the SNP count is truly random and its mean is small, statistical theory suggests we should expect an asymmetric PDF with its mean skewed to the left such as the Poisson PDF, and if the mean count is larger we should expect Normal (aka Gaussian) PDF. Hopefully the empirical data will show us what constitutes “small” and “larger”. And as the number of testers within each sample increases so the associated bar chart should become “smoother”/less “spiky”, i.e. the discrete SNP counts should tend towards a smoother continuous curve, and if not then this could imply that two or more random variables which are not independent of one another are influencing the SNP counts. In such circumstances neither Poisson nor Normal PDFs will be applicable, since both cater only for a single variable, and instead some other PDF might be applicable such as the Log-Normal PDF. This PDF results from the multiplicative product of two or more independent random variables.[63] Such variables might include the size of the population, the size of the family, the age of the fathers, and/or the varying coverage of individual tests – features alluded to in McDonald’s paper.[64] The Gamma PDF might also be relevant.

This hypothetical expectation of the dominant PDF of SNP count samples may be summarised thus:

Mean	expected	< 15 testers	> 15 testers	
SNP count	bar chart	spiky	spiky	smooth
low	skewed to left	modified Poisson	-	Poisson PDF
high	symmetrical	Normal PDF + "t"	-	Normal PDF
any	bi-modal/spiky	Log-Normal PDF		-

We can now turn to the empirical data I have collected in the 17 samples of BigY SNP counts of testers who share a MRCA SNP. My objectives in collecting and analysing this data were:

1. To create a bar chart of the SNP count frequencies and calculate the mean, variance and standard deviation (SD) for each sample.
2. To chi-square test each sample against the Normal, Log_{10}-Normal, Poisson (based on both Mean and Variance, as in practice these input assumptions give different results) and Gamma PDFs, and thus identify the “best fit” PDF for each sample.
3. To determine the 95.4% and 68.3% CIs relevant to these mean SNP counts using three methods:
   (a) CIs derived from the relevant “best fit” PDF. For Normal and Log-Normal PDFs these are the mean +/- SD (for 68.3% CIs) or mean +/- 2*SD (for 95.4% CIs); for the Poisson PDFs I used asymmetric CIs from www.statology.org/poisson-confidence-interval-calculator.
   (b) CIs derived direct from sample data, by manual interpolation of the actual cumulative frequency probabilities of the two tails of each sample.
   (c) CIs derived from the simple, empirically derived formulae developed in section 3.4 above that are readily usable with small samples, and by project admins who are averse to mathematics.

[61] This is even more relevant when seeking the CIs of the product of two PDFs, one or both of which may be asymmetric.

[62] Alternatively a Poisson PDF can be made symmetric by use of a Log-Normal PDF.

[63] Definition from https://en.wikipedia.org/wiki/Log-normal_distribution.

[64] See McDonald 2021, 2-4, 11-12, 23. Other variables might be mutation rates changing over time or the “Founder effect”, or biases in the sample, such as testers being dominantly from USA, and/or having a higher-than-average disposable income, or having been selected because they were close relatives (as opposed to being random self-selected testers). I accept the argument that the relatively large sizes of 18th and 19th century American families may have led to larger SNP counts, but I see no reason why such biases should affect the type of dominant PDF.

Methodology. For each of the 17 samples[65] I have compiled 1 or 2 sheets containing

(1) Sample data (SNP counts & frequencies), & calculations of mean, variance, Standard Deviation;
(2) Log$_{10}$-Normal frequencies, with associated mean, variance, Standard Deviation (SD);
(3) CIs interpolated from sample cumulative frequencies;
(4) Bar chart;
(5) Chi-squared tests against Normal, Log$_{10}$-Normal, Poisson and Gamma PDFs (on Sheet 2 for the larger samples) showing the results of these chi-squared test results in two forms:
(i) the Excel CHISQ.TEST function probability, and (ii) the ∑(actual-predicted)^2/(predicted) ratio;
(6) Summary

I then prepared a Summary (see Appendix B below) of my analyses of these 17 samples, listed in order of the increasing number of testers. The key to columns in this Appendix is thus:

A,B,C: Details of the sample;
D: My source (mostly e-mail references);
E: The number of testers in the sample;
F: top: The mean of the sample; 2nd line: Log$_{10}$-Normal mean;
G: top: Variance; bottom: "smoothness" of the bar chart (subjective);
H: top: SD; 2nd line: Log$_{10}$-Normal SD;
I: mid: α (for Gamma PDFs); bottom: modified Student t factor (for simplified CI estimate);
J: mid: β (for Gamma PDFs); bottom: square root of mean (for simplified CI estimate);
K: 1st five lines: the 5 PDF types against which I have tested each sample, and derived CIs;
Last two lines: 2 additional methods I have used to derive CIs;
L: Excel CHISQ.TEST results copied from individual data sheets (0 = poor fit, 1 = good fit);
M: sums of (actual-predicted)^2/(predicted) similarly copied (low=good fit, high=poor fit);
For L & M **bold font** indicates the PDF that gives the best fit to the sample data;
N: 95.4% CIs calculated from PDF types and from cumulative frequencies;
O: 68.3% CIs calculated from PDF types, from cumulative frequencies, and from simplified estimate;
For N & O shaded background indicates use of PDF best fits inappropriate as they are < [0.80];
bold font indicates cumulative frequency CIs which are wider than the PDF best fit;
italic font indicates the simplified estimate CIs which are narrower than the relevant PDF best fit or the cumulative frequency CIs;
P: cameo of bar chart copied from the individual data sheet.

Problems. In this analysis of the 17 samples I had to address a number of problems:

- The processes of data acquisition and analysis were tedious and error-prone, even when using Excel functions, because of:
  - possible inconsistent determination of callable, phylogenetically-significant SNPs;[66]
  - random errors in manually identifying or counting of these SNPs;
  - possible systemic errors in my analyses (especially in log-normal PDFs & chi-squared tests);
  - random errors in developing Excel data sheets (volume of data makes processing error-prone);
  - misinterpretation of the calculations.
- Neither the Excel CHISQ.TEST critical values nor the ∑(actual-predicted)^2/predicted parameter seem adequate alone to identify the "best fit", and so I took account of both indices.
- While for a "pure" Poisson PDF, the mean is equal to the variance, in practice either can be used to characterise the Poisson PDF against which the sample fit is compared.
- Where the variance-based Poisson PDFs were better (i.e. the R1a and Strother samples) I also tried averaging the mean and variance, but these attempts did not yield better fits.[67]

[65] All the 17 samples are counts of SNPs, private and intermediate, of testers sharing some MRCA SNP. Data sources are cited on the individual data sheets. Note that I normalised the "Royal Stewart" data to a common 10M bps length, and the O'Brian R-DC782 data counts the SNPs back to R-L226, not to the DC782 MRCA.
The data for the Border Irwin R-FGC13746 sample is included as an example in Appendix C. The data sheets for the other 16 samples are available on request.

[66] Note it is the consistency of the SNP counts within each sample that matters, not the consistency between the samples.

- For many samples I was unable to calculate Gamma PDF data, though this doesn't appear critical.
- The optimum method of calculating CIs of Poisson PDFs is contentious: at least 20 different methods can be considered.[68] I have used the asymmetric upper and lower CIs derived from the convenient www.statology.org/poisson-confidence-interval-calculator, which according to www.statology.org/poisson-confidence-interval are $0.5*X^2_{2N,\ \alpha/2}$ and $0.5*X^2_{2(N+1),\ 1-\alpha/2}$ respectively, where X^2 is the chi-squared critical value, N is the number of "Observed Events" (in this application the mean SNP count), and α is the significance level (in this application either 95.4% or 68.3%).
- The CIs derived direct from the sample cumulative frequencies of SNP counts (as opposed to those derived from theoretical PDF models) are sensitive to outliers, have to be interpolated manually, and hence lack reliability; they are inappropriate for 95% CIs and for small samples.

Findings. The following findings emerging from this exercise seem appropriate:[69]

1. Samples of less than about 15 testers sharing some MRCA SNP are of little use in establishing their statistical characteristics.[70] There is thus some merit in developing a simple practical tool, based on analyses of larger samples, for guiding project admins of surname projects with small samples on developing indicative Confidence Intervals for TMRCA estimates based on the convenient mean SNP counts; such a tool may be of some use to admins averse to mathematics.
2. All of 17 samples with more than 15 testers are samples of the larger, real-world populations of the relevant surname/haplotree branch. Each sample is growing over time as more members take NGS tests.
3. Contrary to my expectations, in practice none of the samples show a close fit to their associated Normal, Log_{10}-Normal, Poisson or Gamma PDFs.[71]
4. The best fit I was able to find was with the Border Irwin R-FGC13746 Y700 sample. Ironically here there was little to choose between the closeness for the fits with its associated Normal, Log_{10}-Normal and Poisson (mean) PDFs.
5. Of the 17 samples:
   - 9 have a "best fit" with their Poisson (mean) PDFs.
   - 2 have a "best fit" with their Poisson (variance) PDFs: R1a (poor data, poor fit) and Strother.
   - 4 have a "best fit" with their Normal PDFs: Akins AF06 (data is skewed to the right), O'Brian (ancestral SNP L226 lived over 1,500 years ago), R-L226 Y700 (likewise) and Doherty Y500; in all four their Poisson (mean) PDFs were a "second best fit".
   - 2 have a "best fit" with their Log_{10}-Normal PDFs: R-L513 Y500 and R-L513 Y700; however neither are good fits and both are only marginally better fits than with their Normal PDFs.
6. These findings, albeit inconclusive, can be interpreted several ways. It seems apparent that:
   - for chi-squared and CI calculations based on Poisson PDFs it is preferable to use the mean rather than the variance;[72]

[67] Ralph Taylor found that for the Border Irwin R-FGC13746 data using x=n-2 gave marginally better fits, but my applying similar methods with other samples failed to reproduce any improvements and I have not pursued this sophistication.

[68] See www.ine.pt/revstat/pdf/rs120203.pdf. The Statology algorithm is apparently adapted from the Garwood method.

[69] It is curious that where the mean SNP counts of Y500 and Y700 data can be compared in the samples addressed in Appendix B, the Y700/Y500 ratios are 1.19 (Doherty), 1.15 (Border Irwin), 1.10 (R-L513) and 1.20 (R-L226), whereas the SNP mutation rates of derived from the YDNA Warehouse database have a Y700/Y500 ratio of 1.57.

[70] An extreme example was offered to me by Mary Wiley for eight I-Y20863 BigY500 testers whose SNP counts were: 2: 4; 3: 1; 4: 1: 5:0; 6:2: a "U" shaped distribution rather than the typical "∩" shaped distribution!

[71] I have not attempted fits of Log_e-Normal, Exponential or Binomial PDFs, but the "spikiness" of the samples suggests that close fits with any common PDF are unlikely.

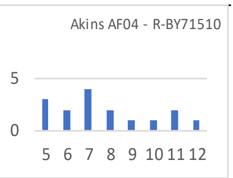
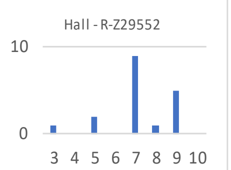
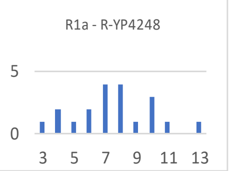
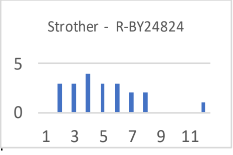
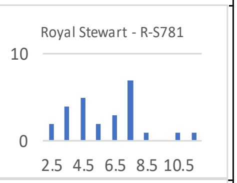
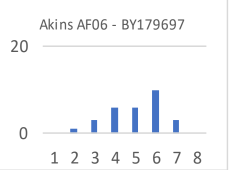
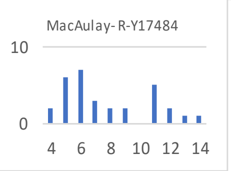
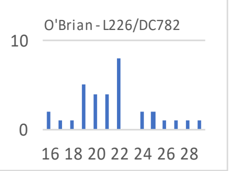
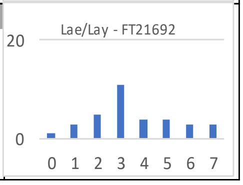
[72] That said, the Poisson PDF is not a "good fit" to every such SNP count. A measure of the relevance of Poisson PDFs to the sample data is the ratio of the variance to the mean, a ratio of 1.00 indicating the sample fits a Poisson PDF exactly:
Variance/mean = 0.3-0.6: Akins (x2), Hall, O'Brian;
Variance/mean = 0.7-0.9: R1a, Royal Stewart, Lae/Lay, Little, Doherty (x2), Border Irwin, R-L513 (x2), R-L226 (x2);
Variance/mean = 1.1-1-2: Strother, MacAulay.

- for samples with mean SNP counts of < c.15, i.e. those for TMRCA calculations within the surname era, the Poisson PDF is the dominant "best fit" PDF;
- for samples with mean SNP counts > c.15, i.e. those used for TMRCA calculations for ethnicity and haplogroup the dominant PDF is less clear.

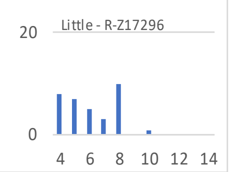
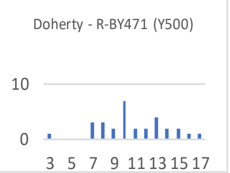
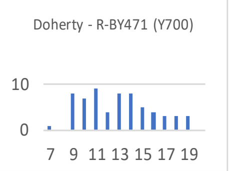
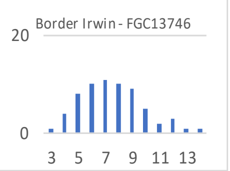
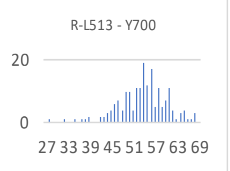
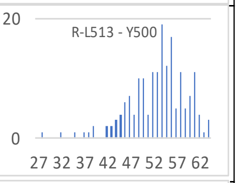
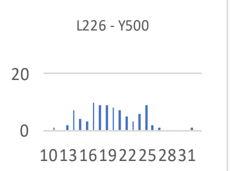
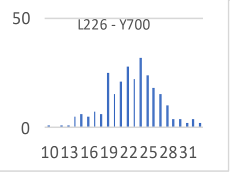
7. The failure of this survey to identify a robust a single "best fit" PDF for all the 17 samples does not invalidate my quest for some indicative CIs relevant to TMRCAs estimated from the convenient mean of SNP counts, even though these TMRCA estimates have been shown to be less accurate than when estimated using the unbiased node-by-node method. As explained above CIs can also be derived be three means: from "best fit" PDFs, from the cumulative frequencies of the sample SNP counts, and by the empirically derived formulae or table developed in section 3.4 above. These CIs are shown in columns N and O of Appendix B, with **bold font** indicating cumulative frequency CIs which are wider than the best fit PDF CIs and *italic font* indicating the empirically derived CIs which are narrower than the relevant best fit PDF CIs or of the cumulative frequency CIs. It will be seen that for curious genealogists, for those averse to lengthy calculations, and for samples of less that c.15 testers, the formulae developed in section 3.4 above offer indicative CIs that are "safe" but conservative.

8. However there is a more fundamental lesson arising from this survey. With the possible exception of the Border Irwin sample, there is no evidence to support my expectations that larger samples (e.g. the R-L513 and R-L226 Y500 samples and the Doherty, R-L513 and R-L226 Y700 samples) are tending towards either a smoother bar chart or a close fit with one of the common PDFs. This conclusion is consistent my belated recognition that while SNP mutations are random, the downstream counts of the SNPs of individuals sharing a MCRA are not random, but are instead biased because of the asymmetric shape of most haplotrees, and possibly other factors such as the impact of changing population size, varying family size, age of the father, Founder effect etc. At present so little is understood about such variables that we cannot even speculate on how they affect SNP counts. These considerations also suggest a Bayesian approach such as that advocated by McDonald, but this is beyond the scope of this paper.

Appendix B - Summary of analyses of 17 samples of SNP counts - 1

A B C D	E	F	G	H	I	J	K	L	M	N	O	P
Data set	No.	Mean	Var-	SD	α	β	PDF type	CHISQ.	∑	Confidence intervals		Bar chart
Ancestral test	of		iance		(Gamma PDFs)		(for chi.sq test & CIs)	TEST	diff.^{2}/f	95.4%	68.3%	
SNP type	testers				TF$_L$	sq.rt.		(Excel)				
source			bar chart			mean						
Akins AF04	16	7.75	4.81	2.19			Normal	0.15	18	3.0 - 12.3	5.3 - 10.0	
R-BY1510		7.45		1.02			Log$_{10}$-Normal	0	3637	5.2 - 9.3	6.2 - 8.3	
Y700							**Poisson on mean**	**0.91**	**7**	3.2 - 15.5	5.0 - 11.6	
							Poisson on variance	0	72			
					12.5	1.61	Gamma					
			fair				interpolated direct			4.1 - 12.7	**4.8** - 11.2	
MC					1.04	2.78	Simplifed estimate				*4.9* - 12.0	
Hall/Smith	18	7.17	2.47	1.57			Normal	0.03	**21**	4.0 - 10.3	5.6 - 8.7	
R-Z29552		6.24		1.06			Log$_{10}$-Normal	0	149	4.1 - 8.4	5.2 - 7.3	
Y700							**Poisson on mean**	**0.08**	29	2.9 - 14.3	4.5 - 11.1	
							Poisson on variance	0	2224			
					20.8	2.90	Gamma	0	-			
			poor				interpolated direct			**2.5** - 9.9	4.9 - 9.5	
DH					1.03	2.75	Simplifed estimate				4.3 - 12.4	
R1a	20	7.55	6.05	2.46			Normal	0	311	2.4 - 12.3	5.1 - 10.0	
R-YP4248		7.08		1.03			Log$_{10}$-Normal	0	huge	5.0 - 10.1	6.0 - 8.1	
Y700							Poisson on mean	0	68			
							Poisson on variance	**0.01**	**27**	3.1 - 15.5	4.8 - 11.4	
					9.43	2.46	Gamma					
			poor				interpolated direct			**2.5** - 13.5	**4.1** - 10.7	
MC					1.03	2.25	Simplifed estimate				*5.2 - 11.0*	
Strother	21	5.05	5.76	2.40			Normal	0.04	22	0.2 - 9.9	2.6 - 7.5	
R-BY24824		4.7		1.02			Log$_{10}$-Normal	0	125	2.7 - 9.9	2.6 - 7.5	
Y700							Poisson on mean	0.33	15			
							Poisson on variance	**0.62**	**11**	1.6 - 11.8	2.9 - 8.5	
							Poisson on mean/var	0.59	11			
					4.42	0.88	Gamma	0	66			
			good				interpolated direct			**1.1 - 17.5**	3.2 - 7.9	
MC					1.03	2.25	Simplifed estimate				2.7 - 9.5	
Royal Stewart	26	5.96	5.17	2.27			Normal	0	775	1.4 - 10.5	3.7 - 8.2	
R-S781		5.52		1.03			Log$_{10}$-Normal	0	61	3.5 - 7.6	4.5 - 6.5	
mixed							**Poisson on mean**	**0.80**	**11**	2.1 - 13.2	3.6 - 9.6	
(normalised							Poisson on variance	0.63	13			
to 10M bps)					6.87	1.15	Gamma	0.15	25			
			poor				interpolated direct			**1.9** - 11.9	**3.0** - 8.3	
IM					1.02	2.44	Simplifed estimate				3.5 - 9.7	
Akins AF06	29	5.03	1.69	1.3			**Normal**	**0.93**	**4**	2.3 - 7.6	3.6 - 5.2	
R-BY179697		4.84		1.02			Log$_{10}$-Normal	0.13	11	2.7 - 6.7	3.7 - 5.7	
Y700							Poisson on mean	0.36	14	1.6 - 11.8	2.9 - 8.5	
							Poisson on variance	0	844			
					15.01	2.98	Gamma		?			
			good				interpolated direct			**1.7 - 7.7**	**3.1 - 6.8**	
MC					1.02	2.24	Simplifed estimate				2.7 - 8.5	
MacAuley	32	*7.59	8.99	3.00			Normal	0.40	19	1.6 - 13.6	4.6 - 10.6	
R-Y17484		7.00		1.04			Log$_{10}$-Normal	-	huge	4.8 - 9.2	5.9 - 8.1	
Y700							**Poisson on mean**	**0.56**	**18**	3.1 - 15.3	4.8 - 11.4	
		*: mean of					Poisson on variance	0.16	29			
		Y29170			6.41	0.85	Gamma	?	114			
		= 11.7	fair				interpolated direct			**1.7** - 14.5	**4.3 - 11.7**	
KM					1.02	2.75	Simplifed estimate				*4.8* - 11.8	
O'Brian DC782	33	21.51	9.76	3.12			**Normal**	**0.88**	**15**	15.2 - 27.8	18.4 - 26.4	
R-L226		21.28		1.00			Log$_{10}$-Normal	0	huge	19.3 - 23.3	20.3 - 22.3	
Y700							Poisson on mean	0.63	22	13.3 - 32.9	16.9 - 27.2	
(filtered							Poisson on variance	0	huge			
by DW)					47.4	2.20	Gamma		?			
			fair				interpolated direct			16.4 - **29.2**	**18.2** - 26.4	
DO/RC					1.02	4.64	Simplifed estimate				16.8 - 28.6	
Lae/Lay	34	3.56	3.25	1.80			Normal	0.785	7	0 - 7.2	1.8 - 5.4	
R-FT21692		2.92		1.16			Log$_{10}$-Normal	0.916	9	0.5 - 5.7	1.8 - 4.1	
Y700							**Poisson on mean**	**0.999**	**6**	0.9 - 9.7	1.8 - 6.6	
							Poisson on variance	0.997	8			
					3.9	1.06	Gamma	?	6			
			good				interpolated direct			**0** - 7.7	**1.3** - 6.1	
MG					1.02	1.89	Simplifed estimate				*1.6 - 6.5*	

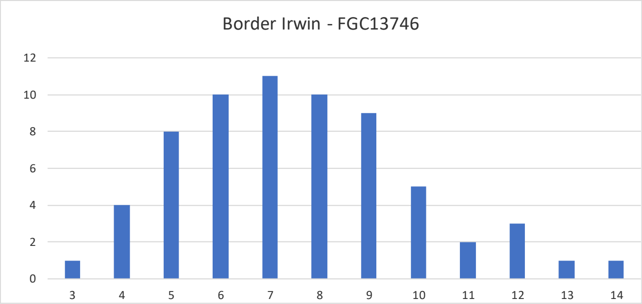
Appendix B - Summary of analyses of 17 samples of SNP counts - 2

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Data set				No.	Mean	Var-	SD	α	β	PDF type	CHISQ.	∑	Confidence intervals		Bar chart
	Ancestral	test		of		iance		(Gamma PDFs)		(for chi.sq test & CIs)	TEST	diff.^{2}/f	95.4%	68.3%	
	SNP	type		testers				TF_L	sq.rt.		(Excel)				
		source				bar chart			mean						
Little				37	5.81	3.78	1.94			Normal	0.15	18	1.9 - 9.7	3.9 - 7.8	
	R-Z17296				5.51		1.04			Log$_{10}$-Normal	0	1904	3.4 - 7.6	4.5 - 6.5	
		Y700								**Poisson on mean**	**0.2**	20	2.0 - 12.9	3.5 - 9.4	
										Poisson on variance	0	115			
								8.94	1.54	Gamma	0	89k			
						poor				interpolated direct			**0.9** - 10.1	**3.3** - 8.4	
		JL/TL						1.02	2.41	Simplifed estimate				3.3 - 9.5	
Doherty				30	10.83	9.41	3.07			**Normal**	**0.90**	**14**	4.7 - 17.0	7.8 - 13.9	
	R-BY471				10.30		1.02			Log$_{10}$-Normal	0	huge	8.3 - 12.3	9.3 - 11.3	
		Y500								Poisson on mean	0.83	16	5.2 - 19.5	7.2 - 15.2	
										Poisson on variance	0.51	19			
								12.5	1.15	Gamma	?	1100			
						fair				interpolated direct			**2.7** - 17.3	7.2 - 14.7	
		ZD						1.02	3.29	Simplifed estimate				*7.5* - 15.9	
Doherty				63	12.90	8.91	2.99			Normal	0.73	18	6.9 - 18.9	9.9 - 15.9	
	R-BY471				12.59		1.01			Log$_{10}$-Normal	0	6651	4.5 - 20.7	8.6 - 16.6	
		Y700								**Poisson on mean**	**0.84**	15	6.7 - 22.3	9.3 - 17.6	
										Poisson on variance	0	265			
								10.7	1.45	Gamma	?0	**9**			
						fair				interpolated direct			8.1 - 19.5	**9.1** - 16.8	
		ZD						1.02	3.59	Simplifed estimate				*9.3* - 18.3	
Border Irwin				65	7.57	5.48	2.34			Normal	0.994	6	2.9 - 12.2	5.2 - 10.9	
	R-FGC13746				7.21		2.39			Log$_{10}$-Normal	0.9994	8	5.3 - 9.0	6.3 - 8.1	
		Y700								**Poisson on mean**	**0.9999**	**5**	3.1 - 15.3	4.9 - 11.5	
										Poisson on variance	0	70			
								10.44	1.38	Gamma	0	907			
						good				interpolated direct			3.1 - 13.5	**4.6** - 10.4	
		JI						1.01	2.75	Simplifed estimate				*4.8* - 11.7	
R-L513				185	53.64	44.63	6.68			Normal	0	199	40.3 - 67.0	47.0 - 60.3	
		Y700			53.21		6.70			**Log$_{10}$-Normal**	0	**163**	41.8 - 68.6	48.5 - 61.9	
										Poisson on mean	0	178	40.0 - 70.3	46.4 - 62.0	
										Poisson on variance	0	52k			
								64.5	1.2	Gamma	-	huge			
						v.spiky				interpolated direct			**37.3** - 66.7	46.7 - 60.6	
		DV						1.00	7.34	Simplifed estimate				46.3 - 64.6	
R-L513				237	48.78	43.16	6.57			Normal	0	90	35.6 - 61.9	42.2 - 55.3	
		Y500			48.31		6.59			**Log$_{10}$-Normal**	0	**83**	35.1 - 61.5	41.7 - 54.9	
										Poisson on mean	0	93	35.8 - 65.6	41.8 - 55.7	
										Poisson on variance	0	330			
								55.2	1.13	Gamma	0	50k			
						v.spiky				interpolated direct			**32.5** - 61.3	42.0 - 55.7	
		DV						1.00	6.98	Simplifed estimate				41.8 - 59.2	
R-L226				87	19.68	15.46	3.94			Normal	0.06	37	11.8 - 27.5	14.8 - 23.6	
		Y500			19.47		3.94			Log$_{10}$-Normal	0.05	40	11.6 - 27.3	15.5 - 23.4	
		(filtered								**Poisson on mean**	**0.66**	**26**	11.8 - 30.7	15.3 - 25.2	
		by DW)								Poisson on variance	0	25k			
								25.1	1.27	Gamma	0	6575			
						fair				interpolated direct			12.5 - 27.0	14.9 - **25.8**	
		DO/RC						1.01	4.35	Simplifed estimate				*15.3* - 26.3	
R-L226				262	23.55	20.85	4.57			**Normal**	**0.38**	**38**	19.0 - 28.1	14.4 - 32.7	
		Y700			25.47		5.37			Log$_{10}$-Normal	0	108	14.7 - 36.2	20.1 - 30.8	
		(filtered				28				Poisson on mean	0.30	39	14.9 - 35.3	18.7 - 29.5	
		by DW)								Poisson on variance	0	231			
								26.6	1.13	Gamma	0	3070			
						good				interpolated direct			**13.6 - 32.8**	17.5 - 27.0	
		DO/RC						1.00	4.85	Simplifed estimate				*18.7* - 30.7	

Appendix C - Border Irwin FGC13746 Y700 SNP counts

(1) Sample data						(2) Log_{10}-Normal calculation					(3) CIs direct from sample data		
c	f	fc	c-mean	$(c\text{-mean})^2$	$f(c\text{-mean})^2$	*$\log_{10} c$*	*$f{*}\log_{10} c$*	*c-mean*	*$(c\text{-mean})^2$*	*$f{*}(c\text{-mean})^2$*	f% of ∑f	cum f%s	
3	1	3.00	-4.57	20.88	20.88	0.477	0.477	0.477	0.228	0.228	1.54%	1.54%	$LCI_{95.4\%}$ @ 2.3% = c. 3.1
4	4	16.00	-3.57	12.74	50.96	0.602	2.408	0.602	0.362	1.450	6.15%	7.69%	
5	8	40.00	-2.57	6.60	52.81	0.699	5.592	0.699	0.489	3.909	12.31%	20.00%	$LCI_{68.4\%}$ @15.6% = c. 4.6
6	10	60.00	-1.57	2.46	24.62	0.778	7.780	0.778	0.605	6.053	15.38%	35.38%	
7	11	77.00	-0.57	0.32	3.56	0.845	9.295	0.845	0.714	7.854	16.92%		
8	10	80.00	0.43	0.19	1.86	0.903	9.030	0.903	0.815	8.154	15.38%		
9	9	81.00	1.43	2.05	18.42	0.954	8.586	0.954	0.910	8.191	13.85%		
10	5	50.00	2.43	5.91	29.54	1.000	5.000	1.000	1.000	5.000	7.69%	18.46%	$UCI_{68.4\%}$ @15.6% = c.10.4
11	2	22.00	3.43	11.77	23.54	1.041	2.082	1.041	1.084	2.167	3.08%	10.77%	
12	3	36.00	4.43	19.63	58.90	1.079	3.237	1.079	1.164	3.493	4.62%	7.69%	
13	1	13.00	5.43	29.49	29.49	1.114	1.114	1.114	1.241	1.241	1.54%	3.08%	$UCI_{95.4\%}$ @ 2.3% = c.13.5
14	1	14.00	6.43	41.35	41.35	1.146	1.146	1.146	1.313	1.313	1.54%	1.54%	
Sum	65	492.00			355.94		55.747			49.053			
Mean = ∑fc/∑f =					7.57					0.858			
										7.21			
Variance = $\sum f(c\text{-mean})^2/\sum f$ =					5.48					0.755			
										5.69			
Normal PDFs:			SD = sqrt.variance =		2.34					2.39			
Gamma PDFs:			Mean = α/β		Var'n = α/β^2								
			α =		10.45								
			β =		1.38								

(4) Bar chart:



(5) Chi^2 tests

		Normal		Poisson		Poisson		Log10-Normal		Gamma	
		mean =	7.57	mean =	7.57	variance=	5.48	mean =	7.21	α =	10.44
		Std. Dev'ı	2.34					Std. Dev	2.39	β =	1.38
c	f, actual	f predicted	$diff^2/f$	f predicted	$diff^2/f$	f predicted	$diff^2/f$	f predicted	$diff^2/f$	f predicted	$diff^2/f$
0	0	*0.08*	0.08	0.03	0.03	0.27	0.27	*0.15*	0.15	0.00	
1	0	0.22	0.22	0.25	0.25	1.49	1.49	0.37	0.37	0.00	
2	0	0.65	0.65	0.96	0.96	4.07	4.07	1.01	1.01	0.00	
3	1	1.65	0.25	2.42	0.84	7.43	5.57	2.30	0.73	0.01	118.77
4	4	3.46	0.08	4.59	0.07	10.18	3.75	4.40	0.04	0.06	255.89
5	8	6.06	0.62	6.94	0.16	11.16	0.90	7.08	0.12	0.24	249.25
6	10	8.85	0.15	8.76	0.18	10.19	0.00	9.54	0.02	0.65	133.52
7	11	10.76	0.01	9.47	0.25	7.98	1.14	10.81	0.00	1.36	68.45
8	10	10.90	0.07	8.97	0.12	5.47	3.76	10.27	0.01	2.32	25.40
9	9	9.19	0.00	7.54	0.28	3.33	9.66	8.20	0.08	3.42	9.11
10	5	6.46	0.33	5.71	0.09	1.82	5.53	5.49	0.04	4.48	0.06
11	2	3.78	0.84	3.93	0.95	0.91	1.31	3.09	0.38	5.34	2.08
12	3	1.85	0.72	2.48	0.11	0.41	16.10	1.46	1.64	5.88	1.41
13	1	0.75	0.08	1.44	0.14	0.17	3.89	0.58	0.31	6.06	4.23
14	1	0.25	2.19	0.78	0.06	0.07	12.67	0.19	3.41	5.91	4.08
15	0	0.07	0.07	0.39	0.39	0.03	0.03	0.05	0.05	5.49	5.49
16	0	0.02	0.02	0.19	0.19	0.01	0.01	0.01	0.01	4.90	4.90
17	0	0.00	0.00	0.08	0.08			0.00	0.00	4.20	4.20
18	0	0.00	0.00	0.03	0.03			0.00	0.00	3.49	3.49
19	0	0.00	0.00	0.01	0.01			0.00		2.82	2.82
20	0	0.00	0.00	0.01	0.01			0.00		*8.37*	*8.37*
∑	65	65.00	6.40	65.00	5.20	65.00	70.16	65.00	8.38	65.00	901.54
Chi^2 value		0.9943		0.99991		0.000		0.9994		0.0675	

(6) Summary

Sample measures:

Mode =	7
Median = (14-3)/2 = 8.5	
Mean =	7.6
Variance =	5.5
SD =	2.3

Comparisons of data with other PDFs:

Best fit: Poisson (mean)
(variance/mean
ratio = 0.7)

Good fits: Log_{10}-Normal
Normal

Poor fits: Poisson (variance)
Gamma

CIs:	95.4%	68.4%
Normal:	2.9-12.2	5.2-10.9
Log_{10}-Normal:	5.3- 9.0	6.3- 8.1
Poisson (mean):	3.1-15.3	4.9-11.5
ex cumulative frequencies:		
	3.1-13.5	4.6-10.4
ex equations (10) & (11):		
	-	4.8-11.7