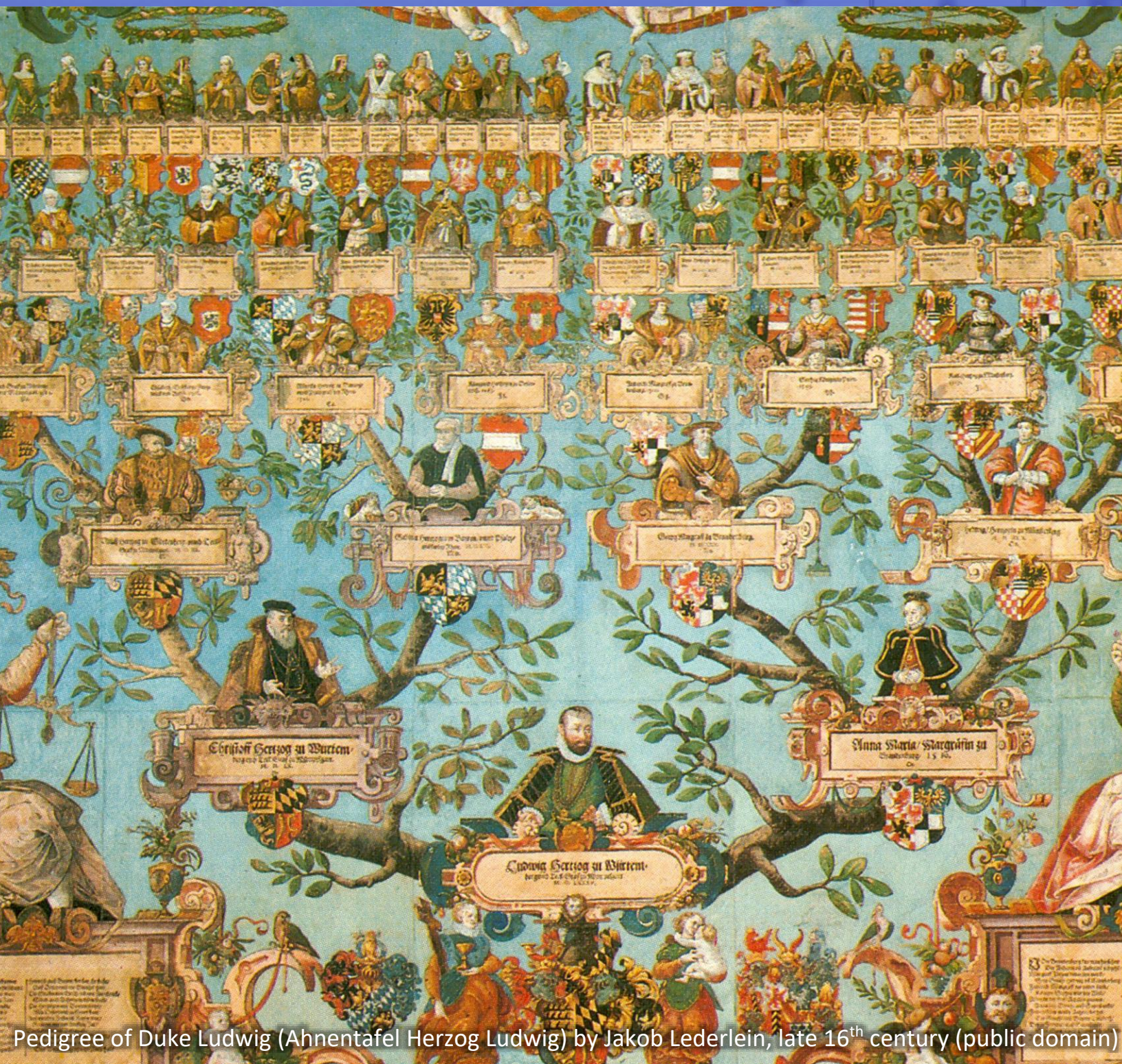
Journal of
Genetic Genealogy
Volume 11 Number 1 (Fall 2023)



Pedigree of Duke Ludwig (Ahnentafel Herzog Ludwig) by Jakob Lederlein, late 16th century (public domain)

Journal of Genetic Genealogy

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

Journal of Genetic Genealogy

Taking Account of Human Genetics in Identity Research

Introduction

It is sometimes remarked that the quest for identity is the real aim of genealogy. But what is identity; what does human genetics offer to the identity debate; and how does genetic genealogy impact on cognitive views of identity? Currently, mention of DNA tends to polarize rather than unite. For example, Brodwin (2002; pp. 323, 324) opines that … *to the dismay of anthropologists who fancy themselves as the cultural avant-garde, essentialist identities grow ever more powerful and seductive*. He takes the view that emerging knowledge in the field *… adds the cachet of objective science to the notion that one's identity is an inborn, natural, and unalterable quality. Rapid advances in sequencing and analyzing the human genome have strengthened essentialist thinking about identity ...* He adds *… Emerging genetic knowledge thus has the potential to transform contemporary notions of social coherence and group identity*. Comfort (2019; p. 170), on the other hand, remarks that [d]*efining the self only in biological terms tends to obscure other forms of identity, such as one's labour or social role*. Hauskeller (2004; p. 296) concurs and states: *Although DNA is no doubt real, it is clear that there are systematic problems in employing it for purposes such as establishing identity*. Clearly, the use of human genetics in identity is a controversial topic.

The most comprehensive article that I have come across about how genetics relates to identity shows that there is little agreement, even among scientists. The paper in question is a systematic scoping review of the concept of 'genetic identity' by Goekoop, *et al* (2020; pp. 1, 16). The authors found that: *… a clear understanding of the term is lacking … Overall, the diversity in the use of 'genetic identity' in the reviewed literature demonstrates that the term is used differently in different contexts, but also within each context the meaning of the term can vary widely* [and be used] … *in contradicting ways* (Goekoop, *et al*, 2020; pp. 1, 16).

The present paper provides a fresh look at this complex subject, taking care to define terms clearly and avoid—in as far as possible—the most obvious contradictions. It does not consider *artificial* interventions in human genetics. In other words, no medical therapies or interventions like organ transplants or blood transfusions are contemplated when considering biological identity. At this juncture, determining when a significant rupture has been caused and the biological identity of the individual has been compromised is open to interpretation and will be influenced by growing information about the functioning of our genome. However, this discussion is outside the remit of the presentation. In addition, the focus of the paper is on nuclear DNA rather than on mitochondria.

Hauskeller points out that *DNA is real.* But how do we account for it or incorporate it into wider identity studies? How does genetics impact on our cognitive or subjective views of identity which currently predominate? Intuitive approaches may appear more reasonable and compelling to many, than relying on what seems to be an indifferent chemical molecule to define the term. But surely DNA has a complementary role to play? It is only by knowing more about the subject that identity research will mature and answer questions about the role of human genetics in determining who we are. The purpose of this paper is to: 1) Provide a brief introduction to DNA and propose the human genome as **a** basis for genetic or biological identity; 2) Show how studies in human genetics and cognition can be compatible and complementary rather than irreconcilable or contradictory; and 3) Begin a discussion about how to capture mutually reinforcing benefits by amalgamating human genetics and subjective studies of cognition. The wider focus of the paper is on introducing biological or genomic identity as a base for identity, in contrast to cognitive perspectives, which are recognised here as expressions of how we identify

or are identified.

'Sameness'

The confusion associated with 'identity' can be traced to its late 16th century origins, combined with a conundrum that stretches into the remote past. The word emerged from Medieval Latin *identitas* or *idem* meaning 'sameness' or 'same'. Unfortunately, 'sameness' has been difficult to comprehend. This is well illustrated in the ancient Greek paradox of the Ship of Theseus, which was kept as a monument to its hero at Athens. As time passed the planks had to be changed one by one when they began to rot. At what point, philosophers asked, did the original ship cease to be itself? If every plank, rib and panel were replaced, along with the nails that held them together, how could it be the same ship? Surely it was only a replica. We know that the chemical constituents of our bodies are continually changing. As such, are we the same person as we move through time? In sum, there have been no adequate answers and thus 'identity' has suffered a crisis of meaning. Fortunately, the 'sameness' paradox is no longer a stumbling block in biological terms, at least in relation to the question: Are we always the same (physical) person?

Personal identity

Identity has an individual and a group dimension. Since 1987, thousands of forensic cases have been decided with the assistance of genetic fingerprinting (Roewer, 2013). Confidence about its application is generally based on the observations that our individual DNA sequences are unique. Besides, observations indicate that the nuclear DNA of every cell remains the same throughout the life of the organism. In other words, science reveals what is meant by 'sameness' in biological terms in who we are. If this is correct, then our DNA sequences provide a basis of our genetic identity at the personal level, which can be defined using older and more rigorous meanings of the word. For example, the Oxford English Dictionary (OED 2nd Edition, 1989), defined personal identity as: *The sameness of a person at all times or in all circumstances; the condition or fact that a person is itself and not something else*. It is the definition that is applied in this paper.

However, Dr. Alexander Hoischen of Radboud University Medical Center in the Netherlands, who was part of a research project into mosaic mutations stated that: *The textbook knowledge that our genome is identical in all the cells of our body is probably not true* (Azvolinsky, 2015). As sequencing practices are becoming more widely used and as techniques are improving, an increasing number of anomalies are being recorded. Despite advances, limitations of current technologies still hamper our understanding of the extent of these changes (Acuna-Hidalgo, *et al*, 2015; p. 67). Mutations may occur in early development at the zygote stage, but they may also appear in post-zygotic development. Account must also be taken of fetomaternal microchimerism which is a research field in its infancy and which refers to the bi-directional exchange of a small number of cells and cell-free fetal DNA through the placenta (Murrieta-Coxca, *et al,* 2022 and Rosner, *et al*, 2021). Although our knowledge is in its infancy, the exceptional intrusion of complex DNA appears to have important implications during and after pregnancy (Rosner *et al*, 2021). The repercussions of fetal DNA transfer and post-zygotic mutations on the definition of personal identity require careful consideration. Other considerations are the mutations that occur in a stem cell whose descendants become a specialised organ. The tissue of such an organ will appear different compared to all other parts of the body with respect to that mutation. The term for this anomaly is a 'mosaic mutation'. Anomalies can also arise, caused by the fusion of two zygotes during the early embryonic stage, forming a *fusion chimera* (Madan, 2020). The appearance of such differences has given rise to the suggestion that some of our cells carry different versions of our genomes (Azvolinsky, 2015) or that some people carry multiple genomes (Ledford, 2019; Lupski, 2013). However, for the purposes of the

present discussion, the genome is understood to mean *the complete set of genetic material present in an organism*, whatever its complexities. Therefore, all material—whether original, mutation, or even blended genotypes, as in the case of fusion or fetal chimeras—is included in the term and this is supported by the arguments that the personal genome is singular in expressing its nature and, furthermore, that the entire organism is a viable unit.

We know that mutations are relatively low in the human body. For example, in each new generation *one error in every 10^8 base pairs gives rise to 30–100 genome-wide de novo or new mutations* (Acuna-Hidalgo, *et al*, 2015; p. 67). This means that 0.000065 percent of base pairs are changed, leaving 99.999935 percent unchanged. Every mutation affects a minor part of the entire genome. In other words, a high level of structural conservation of nuclear DNA is retained in the bulk of our underlying genetic sequences across the whole organism. Nonetheless, the contribution of the post-zygotic mutation rate is unknown. New studies continue to emerge. One result reveals that early mutations were estimated to be 0.34×10^{-8} for one of a monozygotic twin pair and 0.04×10^{-8} for the other (Dal, *et al*, 2014). This shows that so-called 'identical twins' are not identical. Another study detected mutations among multiple samples obtained from the same individuals that included samples of blood, saliva, hair follicle, lining of the cheeks, urine, and semen. Results indicate that a clinically unremarkable person might harbour post-zygotic mutations corresponding to around 1.5×10^{-8}-4.4×10^{-7} per nucleotide per individual (Huang, *et al*, 2014; p. 1319). These are relatively low numbers and confirm the structural conservation of the bulk of our genome. It may seem reasonable, therefore, to base 'sameness' of the person on the observation that most of our genome remains the same. But no matter how small the modification is—it *is* a change, and the strict meaning of 'sameness' is infringed. This scrupulous attention to detail is important. After all, if a small mutation gives rise to a gene, or influences a group of genes that induce susceptibility to cancer or another malady, it is highly significant for the whole organism. So, even though the rates of change are relatively small, the implication that the genome is not the same *at all times or in all circumstances* in terms of its structural makeup is irrefutable in these instances. As mutations are part of the life of our genome, DNA sequences—alone—cannot be used to define personal identity ('sameness') in all cases. For this reason, the term must be qualified. The question we must now answer is: How do we know that we are the same person if post-zygotic mutations take place? In what way are we the same under these circumstances? It can be demonstrated that 'sameness'—though not contained in all DNA sequences—is present in the individual's genome, which conforms to a specific set of unchanging conditions. These include:

1. Continuity: development from its beginning to end-of-life is an uninterrupted series of events (in terms of life processes or metabolism)
2. Viability: the genome survives as a single entity
3. Individuality: the personal genome is singular in expressing its nature; and
4. Uniqueness: it is unlike any other genome

Continuity intimately links subsequent post-zygotic mutations to the same genome. A separate personal genome is not created because mutations take place. Mutations are part of the nature of many—if not most—individual genomes. Not only is the bulk of the underlying genome conserved in terms of its structure, it remains viable at all times. Viability is an important quality in that it overcomes the suggestion that the chimera or mosaic genome is, in fact, two or more people in one. A person cannot be divided up—particularly if vital organs are involved—and remain viable. And individuality is maintained from the point of fertilization until death. Furthermore, individuality is highlighted by the uniqueness of the makeup of the genome itself, which must also take account of changes in epigenetics (i.e., different genes being 'turned on' or 'off') which cause genomes to differ. On a separate issue: even monozygotic twins (or multiples) share a common beginning when the parental gametes unite. However, viability and continuity

in terms of life processes or metabolism, are not compromised during the shared stage for each genome. The case of conjoined twins, especially those sharing vital organs, represents an exception and requires further consideration.

The possibility of infringing 'uniqueness' where cloning has occurred is a perennial question, assuming that it was ethically permissible, which it is not. However, even in a case in which the nuclei of both donor and clone are the same, DNA reacts in conjunction with the cellular fluid (cytoplasm) in which it is embedded. Therefore, to be the same the nucleus of the clone would have to be introduced into the same host as that of the donor. Many discrepancies would arise as a result of epigenetic changes or gene expression. The developing cells would also be subject to random mutations and the effects of fetomaternal microchimerism. Furthermore, even a clone of oneself would differ from the donor in that its subsequent development would never replicate the original.

Another issue that is sometimes raised in relation to biological or scientific identity is the complexity of new information emerging from microbiology which forces us to take into consideration the relationship between the host and the microbiome (Liu, *et al*, 2021) especially in relation to symbiosis and pathogenesis. Comfort (2019; p. 169) goes as far as suggesting that the 'biological self' has been reframed as a cluster of communities rather than the individual. However, from the point of view of human identity, the integrity of host and microbial communities are maintained. Symbiosis or virulency are, after all, broad terms to describe *different* organisms interacting with the host (Eloe-Fadrosh and Rasko (2013) and Méthot and Alizon (2014)). Besides, a person's immune system is unique with regard to other living organisms and even when compared to other individuals of the human species, including a monozygotic twin. Consequently, individuality is manifest in the boundaries drawn by one's immune system (Pradeu, 2012; pp. 7, 8).

Communal identity

According to Keogh (2019; p. 17): *In biological terms, a species is generally defined as a group of organisms capable of successfully exchanging genes, or in other words, capable of interbreeding to produce fertile offspring. Ability to reproduce with our own kind is, therefore, the essential constant that distinguishes us as human. This means that communal identity becomes manifest through the 'trinitarian' act of reproduction, in which two personal identities of the opposite sex give rise to a new personal identity.* The main question to answer here is: How does communal identity 'become manifest' through procreation? Focus is placed on the formation of the individual genome irrespective if the parents produce other offspring at another time or several unique genomes through the one reproductive event. The instant that gene fusion takes place, which describes the essence of reproduction, is expressed by Condic (2008; p. 3): *Following the binding of sperm and egg to each other, the membranes of these two cells fuse, creating in this instant a single hybrid cell: the zygote or one-cell embryo … Cell fusion is a well studied and very rapid event, occurring in less than a second. Because the zygote arises from the fusion of two different cells, it contains all the components of both sperm and egg, and therefore the zygote has a unique molecular composition that is distinct from either gamete. … These modifications block sperm binding to the cell surface and prevent further intrusion of additional spermatozoa on the unfolding process of development. Thus, the zygote acts immediately and specifically to antagonize the function of the gametes from which it is derived … Clearly, then, the prior trajectories of sperm and egg have been abandoned, and a new developmental trajectory—that of the zygote—has taken their place*. Clearly, the essence of reproduction occurs … *in less than a second* and, although communal identity becomes manifest through it, the underlying 'sameness' of the group is ongoing and not confined to a single instant. How, then, is this 'sameness' or identity to be understood?

The collective dimension in procreation is underpinned, firstly, by the degree of randomness associated with the parental encounter, which involves the union of two representatives of the opposite sex from the segment of the genomic spectrum that is potentially fertile at that moment. Even though panmixia is only theoretically feasible, the entire component of humanity that falls into this category, is open, at least in biological terms, to the possibility of interbreeding. Random events make potentialities concrete. Irrespective of the capacity—or incapacity—of offspring to procreate, all humans are connected to the reproductive process in that they have been created by it.

Turning to the wider picture: DNA patterns do not correlate well with the normal divisions of humanity like race, ethnicity, culture, or nation. This does not rule out genetic patterns that are more common in certain groups. However, ethnicity and race are shown to be non-genomic (Kim, *et al*, 2023). Genetics is the new classifier and shows that it is not possible to recognise these categories in terms of our DNA because individual genomes are part of a virtual continuum of genetic variation around the world (Marshall, 1998). The interconnectedness of modern humanity resulted from our origins in a small bottleneck of people in Africa—our *ancestral singularity*—some 200,000 years ago and subsequent developments (Keogh, 2016; pp. 146 ff). Even the probable interbreeding with Neanderthals and Denisovans failed to establish a separate species and inhibit intermixing. From a genealogical viewpoint, an individual is composed of multi-dimensions in their many genetic pathways out of the past. In other words, genomic composition is made up of DNA mixes that often emerge from sub-groups that intertwined uneasily in society. However, knowledge that we share a common biological identity helps alleviate divisiveness in terms of logic, but often requires a long process of emotional reconciliation before a satisfactory acceptance is established between contested components in our makeup.

The entire human genome, which is composed of unique and therefore different individual genomes, constitutes the complete spectrum or range of DNA sequences that exist. It follows that no human sub-groups—no matter how large or small—are made up of individuals who share a 'sameness' that is exclusive to that cohort. The communality within sub-sets that we observe across this spectrum contain similarities that may be helpful in the study of human behaviour, but they do not constitute separate identities. Resistance to the notion of intrinsic internal divisions in the human family—beyond the individual—means that communal identity is located exclusively at the universal group level, which is composed of all living individuals or personal genomes. The *nature* of the characteristics that imbue the individual with personal identity (continuity, viability, individuality and uniqueness) are shared by parents and offspring. In like manner, this nature is shared by all people and provides us with a basis for expressing communal identity or 'sameness' at the universal level. Consequently, collective, universal or biological identity may be defined as: *The sameness of the group at all times or in all circumstances; the condition or fact that the group is itself and not something else.*

Power of DNA

The founder of the modern science of genetics was Gregor Mendel who discovered the principles of biological inheritance between 1856 and 1863. But the field only evolved rapidly after the derivation of the structure of the double helix in mid-20th century (Pray, 2008). Another significant recent advance was made by sequencing the entire human genome in 2003. Our rapidly growing knowledge has initiated what is popularly called *the DNA revolution*. The revolution is advancing through the increase in our understanding of how DNA works and how it can be manipulated, which is not without risks. We now know that DNA contains the biological instructions that make each species unique. It is not a passive molecule; *it is highly dynamic, enwrapping layers of complexity … Its function as a universal genetic*

material is among the most highly conserved qualities of living things. Its system of four bases, when overlaid with spatial and temporal controls, governs biology across the entire scale of life and actively contributes to all living processes (Duzdevich *et al*, 2014; p. 3072). It contains the instructions needed for an organism to develop, survive, thrive, reproduce and pass on genetic information from one generation to the next (hereditary). The roles played by nature versus nurture or—more likely—the complex interactions between the environment and the way DNA assuages its impact to give the organism the best chance to stay alive and prosper, must be considered. It is also known that DNA influences personal attributes. Plomin (2019; p. viii) states that he is *not aware of a single psychological trait that shows no genetic influence*. Duncan *et al* (2019; p. 1518) reinforce the link between disease and genetics when they point out that *all major psychiatric disorders have now been shown to be polygenic*. Furthermore, specific loci have been identified through genome-wide association studies (GWAS) rather than by way of specific candidate genes. However, these authors also note that this approach requires *massive increases in sample sizes (i.e. encompassing tens of thousands of participants, or more)*. The advantage of very large sample sizes is that they tend to avoid the dangers of non-repeatability in experimental outcomes, which is a stumbling block to scientific progress, especially considering the current crisis of replicability (Pashler and Harris, 2012). [U]*ntil now, psychologists have had to rely on behavioural symptoms to diagnose disorders*. Genetics is beginning to offer *a causal basis for predicting disorders rather than waiting until symptoms appear and then trying to use these symptoms to diagnose disorders* (Plomin, 2019; p. 66). These results are likely to improve as more GWAS—using very large samples—become available.

However, a central role for DNA in personal and communal identity tends to trigger antagonism based on concerns that it is a form of essentialism that equates with **genetic** determinism or reductionism. And these concerns are intensified by the assumption that human behaviour is, therefore, under the full control of an individual's genes at the expense of the roles of the environment, learning or free will. One way to dispel the notion of genetic absolutism is to consider a scholar who becomes convinced by an argument and is converted to a new way of thinking and acting after reading a ground-breaking scientific paper or book. Genetics allows the subject to see the hieroglyphs on the page, translate them into meaningful thoughts, commit them to memory and compare them to knowledge already accumulated, thereby forging associations, stimulating comprehension, and coming to conclusions that are independent of the personal genome. The incoming words and ideas are the principal agents activating segments of the brain during this process, not the underlying DNA. Julian Baggini (2015), reflecting on human genetics, comes to a nuanced conclusion, which discounts an 'either/or' argument in which either genetic absolutism, the environment or unrestricted human freedom rule supreme. A balance between these aspects reflects the human condition more accurately. However, historically, we are in the difficult process of determining the limits of these factors. On the other hand, 'scientism'—or the belief that material science is the only source of authentic knowledge—is unsound and may be discounted.

Discussion

Most of this paper has been dedicated to introducing the genomic base for genetic or biological identity. In what way might the approach impact on studies in the humanities that rely mainly on subjective cognitive analysis to define and understand identity? Modern classical thinkers and philosophers tended to avoid answering the question 'What is identity?' in terms of material content and, instead, sought to find 'sameness' in intellectual traits. In simple terms, the main question they tried, and failed to answer, is: Does our 'sameness' persist over time; and do we possess an unchanging 'essence'? Hume attempted to overcome the conundrum by suggesting that sameness resides in 'memory'; Descartes favoured 'thinking'; Locke opted for 'consciousness'; Erikson advocated 'ego'; while 'sense-of-self' gained popularity among a wide spectrum of social scientists. The problem in these cases is: How to account for

persistence when we sleep, or if we happen to be in a coma, or if we have the misfortune of getting Alzheimer's Disease? And how do we account for persistence before we had a remembered memory of any sort, as in our time in the womb? On the other hand, does our sense-of-self not change over time? What about personality? **Surely** that, too, can change. Personality modifications may occur when certain physical changes are induced in the brain. Considering these objections, many philosophers began to avoid the suggestion that we have an unchanging essence that makes us who we are. The reason for the turnaround is based on the conviction that 'sameness' is difficult to prove, and therefore, is unsolvable. But failure to reach a satisfactory conclusion exacerbated the problem and deepened the 'identity crisis' as it morphed into what have been called postmodern and constructivist strands of academic thought. These trends began in the 20th century and have made the topic more puzzling by sharing … *a fundamental disbelief in the existence of any 'objective truth' and 'definiteness' – conceptual or otherwise* (Prusch, 2017; p. 10). Consequently, identity is portrayed as not being static, is subject to fluxes and does not exhibit a *sensu stricto.* This inference *… provides the basis for understanding that people's identities may have many different facets, can change all the time and might even contain contradictions*.

However—echoing Hauskeller's observations—*DNA is no doubt real*. It is unwise, therefore, for social scientists to ignore scientific observations in their research. One possible way forward is for the academic world to acknowledge that revelations in biology point to an underlying physical sameness in the individual and in the community. Given that this approach can be demonstrated independently and repeatedly through empirical means allows it to be applied to research in social science and studies of the modern classical thinkers and philosophers, as well as genealogists. The collective reasoning of these specialists provides intermittent glimpses of continuity of the person through memory, thinking, consciousness, ego, narrative and sense of self. The principles on which biological identity are based offer an underlying framework to affix the views of all these scholars and practitioners, thus strengthening their theories. However, 'sameness at all times' transfers to the material content of the personal genome. This change allows sociologists to advance with greater ease in new directions because they are unhindered by the need to reify identity from subject notions of self. Clearly, studies in human genetics and subjective cognition, are compatible and complementary rather than irreconcilable or contradictory. The amalgamation of biological and subjective views will—I believe—also help to determine when a significant rupture has been caused in biological identity through artificial interventions that interfere permanently with or eliminate cognitive continuity. However, clarifying this and many other aspects of identity lies well into the future. To have any chance of success there is an urgent need to begin a serious discussion about the merger of empirical and subjective studies in identity so that the mutually reinforcing benefits of both human genetics and subjective cognition are maximised.

Conclusions

The reluctance to engage in human genetics was highlighted in a presentation I made at the 29th Annual Conference of the International Society for Research on Identity (ISRI) at Tufts University in Boston in 2023. In a rapid and informal oral response from my audience, around 80% were sceptical, had doubts, or knew little about the application of human genetics in identity studies. One participant mentioned its misuse in the past. The varied responses reflect the ongoing ethical and social discussions that often accompany advancements in genetic research and its potential implications. However, an African American participant was enthusiastic about the use of genetics, citing its value as a method of tracing origins in Africa for a people whose history was denied to them due to the slave trade. Her comments echo a wider reality. Genetic genealogists and many family historians have little difficulties in fusing DNA testing with traditional genealogy, oral history, and documentary records, because it has proven to be a powerful research tool. It offers the potential to overcome the challenges and obstacles that often arise

when tracing ancestry through purely non-genetic methods. However, like social scientists, genealogists often view cognitive perspectives as 'identity', which can cause confusion, particularly if identity is viewed as unchanging. Tests results have often produced surprising results, forcing practitioners to change their views of self. It would, however, be incorrect to suggest that their underlying genetic identity has changed; they have discovered something extra about it. For simplicity the dilemma is avoided if cognitive perspectives, which are subject to change, are recognised as expressions of how we identify or are identified, rather than our 'identity'. Viewed from this perspective, genetic genealogy may be seen as a forerunner or vanguard in that it has the potential to merge biological and cognitive perceptions on identity, leading to a more comprehensive understanding of who we are and where we come from, with wider implications for reshaping our understanding of identity itself.

Author: Raymond M. Keogh

Affiliation: Director, Our Own Identity (M. Ag. Sc. (for.) University College Dublin)
Email addresses ourownidentity@gmail.com or raymond@ourownidentity.com

References

Acuna-Hidalgo, R., Bo, T., Kwint, M. P., van de Vorst, M., Pinelli, M., Veltman, J. A., Hoischen, A., Vissers, L. E. L. M., and Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. The American Journal of Human Genetics. Vol. 97 (1): 67-74.

Azvolinsky, A. (2015). Mosaic Mutations May Not Be Rare. TheScientist.

Baggini, J. (2015). Do your genes determine your entire life? The Guardian. (19th March).

Brodwin, P. (2002). Genetics, Identity, and the Anthropology of Essentialism. Anthropological Quarterly. Vol. 75 (2): 323-330.

Comfort, N. (2019). How science has shifted our sense of identity. Nature. Vol. 574; 167-170.

Condic, M. L. (2008). When does Human Life Begin? A Scientific Perspective. Westchester Institute White Paper Series. The Westchester Institute for Ethics and the Human Person, Thornwood, NY.

Dal, G. M., Ergüner, B., Sağıroğlu, M. S., Yüksel, B., Onat, O. E., Alkan, C., and Özçelik, T. (2014). Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. Journal of Medical Genetics. Vol. 51 (7): 455-459.

Duncan, L.E., Ostacher, M. and Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. Neuropsychopharmacol. Vol. 44; 1518–1523.

Duzdevich, D., Redding, S., and Greene, E. C. (2014). DNA dynamics and single-molecule biology. Chemical reviews. Vol. 114 (6): 3072–3086.

Eloe-Fadrosh, E. A., & Rasko, D. A. (2013). The human microbiome: from symbiosis to pathogenesis. Annual review of medicine, Vol. 64; 145–163.

Goekoop, F. M., van El, C. G., Widdershoven, G. A. M., Dzinalija, N., Cornel, M. C., and Evans, N. (2020).

Systematic scoping review of the concept of 'genetic identity' and its relevance for germline modification. PLOS ONE. Vol. 15 (1), e0228263.

Hauskeller, C. (2004). Genes, genomes and identity. Projections on matter. New Genetics and Society, Vol. 23 (3): 285-299.

Huang, A. Y., Xu, X., Ye, A. Y., Wu, Q., Yan, L., Zhao, B., ... & Wei, L. (2014*).* Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. Cell Research. Vol. 24 (11): 1311-1327.

Keogh, R. M. (2016). Shelter and Shadows. Our Own Identity. Ireland.

Keogh, R. M. (2019). DNA & the Identity Crisis. Philosophy Now; Issue 133; 16-17.

Kim, B-J., Choi, J. & Kim, S-H. On whole-genome demography of world's ethnic groups and individual genomic identity. Sci Rep 13, 6316 (2023).

Ledford, H. (2019). The human body is a mosaic of different genomes. Nature News (6 June).

Liu, X., Tang, S., Zhong, H. *et al* (2021). A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. Cell Discovery. Vol. 7 (9).

Lupski, J. R. (2013). Genome Mosaicism – One Human, Multiple Genomes. Science. Vol. 341 (6144): 358-359.

Madan, K. 2020. Natural human chimeras: A review. European Journal of Medical Genetics. Vol. 63, Issue 9.

Marshall, E. (1998). DNA Studies Challenge the Meaning of Race. Science. Vol. 282 (5389); 654-655.

Méthot, P-O. and Alizon, S. (2014). What is a pathogen? Towards a process view of host-parasite interactions. Virulence. Vol. 5 (8): 775-785.

Murrieta-Coxca, J. M., *et al* (2022). Addressing microchimerism in pregnancy by ex vivo human placenta perfusion. Placenta. Vol. 117; 78-86.

Pashler, H. and Harris, C. H. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. Perspectives on Psychological Science. Vol. 7 (6): 531-536.

Plomin, R. (2019). Blueprint – How DNA makes us who we are. Penguin.

Pradeu, T., 2012. The Limits of the Self. Immunology and Biological Identity. Oxford University Press; pp. 7, 8.

Pray, L. (2008). Discovery of DNA Structure and Function: Watson and Crick. Nature Education. Vol. 1 (1): 100.

Prutsch, M. J. (2017). Research for CULT Committee – European Identity. European Parliament, Policy Department for Structural and Cohesion Policies, Brussels.

Roewer, L. (2013). DNA fingerprinting in forensics: past, present, future. Investigative Genetics. Vol. 4 (22): 1-10.

Rosner, M., Kolbe, T. and Hengstschläger, M. (2021). Fetomaternal microchimerism and genetic diagnosis: On the origins of fetal cells and cell-free fetal DNA in the pregnant woman. Mutation Research-Reviews in Mutation Research. Vol. 788 (108399): 1-11.

Journal of Genetic Genealogy

CASE STUDIES USING GENEALOGICAL JUNCTIONS AMONG UNPLACED DNA MATCHES TO IDENTIFY UNKNOWN ANCESTORS[1]

By Lars-Gunnar Lundh
Professor emeritus, Department of Psychology, Lund University, Sweden

Email: Lars-Gunnar.Lundh@psy.lu.se
ORCID: https://orcid.org/0000-0002-1649-969X

Abstract

The present paper describes case studies where the purpose was to identify two unknown ancestors: the author's paternal grandfather [FF] and his paternal grandmother's maternal grandfather's father [FMMFF]. A four-step method was used: (1) The author's and his brother's closest atDNA matches were identified, and the genealogical relationships to these matches were searched for by a comparison with their family trees. Those atDNA matches for which no explanatory genealogical relationship could be found were named *unplaced* DNA matches. Nineteen unplaced DNA matches were found who shared more than 100 cM with the two brothers. (2) The family trees of these unplaced DNA relatives were compared to each other, and common ancestors who recurred in the family trees of at least two unplaced DNA relatives were referred to as *genealogical junctions*. The four "strongest" genealogical junctions were focused (involving, eight, four, three, and three unplaced DNA matches, respectively). (3) An analysis of the *connectedness* between these genealogical junctions (e.g., marriages between descendants) were used to generate two hypotheses: the FF Norberg hypothesis, and the FMMFF Jon Pehrsson hypothesis. (4) The FF Norberg hypothesis was tested in collaboration with four descendants of the Norberg family who tested their atDNA, and one of them who also tested his Y-DNA, and the results fully supported the hypothesis. The FMMFF Jon Pehrsson hypothesis was tested by segment triangulation methods, to see if DNA segments shared with descendants of Jon Pehrsson's (*1795) children overlapped with DNA segments shared with descendants of the author's FMMF. Support for such DNA sharing was found on seventeen of the chromosomes, and the most conclusive support was found on chromosome 16. The results are discussed in terms of the distinction between proofs and evidence, the sensitivity of the genealogical junctions method, the importance of having one's siblings DNA tested, the confounding factor of the number of descendants of a hypothesized ancestor, and the choice of cut-off for identifying the set of unplaced DNA matches.

Key words: unknown ancestors, atDNA, Y-DNA, segment triangulation, genealogical junctions

[1] I wish to thank Sven-Erik Johansson and Robert Eckeryd for invaluable help during the years that I have been working on this project. Without having access to Sven-Erik Johansson's carefully organized genealogical database *Kråken.se*, I would probably never have embarked on the project. And if I had, it would most certainly have taken me much longer to complete. Sven-Erik Johansson has also been helpful in answering all kinds of questions during the process. Without Robert Eckeryd's doctoral thesis on unmarried mothers in this part of Sweden during the relevant periods, and his generous sharing of data about my paternal grandmother collected during his doctoral work, crucial information about her had been missing. I am also very grateful to my brother and my newly discovered cousins for their willingness to contribute to the process by testing their DNA; without their help this study could not have been completed. Finally, I want to thank anonymous reviewers for very valuable comments on a previous version of the manuscript.

1. INTRODUCTION

The present author's father (here referred to as F) was born in 1913 in Nordmaling parish, Västerbotten county, in northern Sweden. His mother Nanny Elina Ekholm (here referred to as FM) was unmarried and already had one child, a daughter born in 1910. She worked as a milkmaid on a farm, and as she found it difficult to take care of two small children, she left F in a foster family where he was raised. According to the birth records F's father (here referred to as FF) was unknown.

There had been many ideas over the years of who FF could be. For example, some had guessed that FF was a father or a son in the foster family. Another lead was followed by the municipality's poor welfare board when they initiated a paternity trial in 1916 against a sawmill worker (here referred to as JL) who had once claimed to be the father. The municipality's poor welfare board made a monthly contribution to the foster family for their expenses raising F, and now they wanted the biological father to take his responsibility and pay. During the trial, however, JL retracted his confession and denied ever having been sexually involved with FM.

During the paternity trial, which was extended over more than three years (as JL repeatedly failed to attend the court hearings), FM got married and moved to another parish together with her daughter. Her new husband, who was a widower with four children of his own, did not want to take responsibility for raising F, who stayed with his foster family. As has been documented by Eckeryd (2017) in his doctoral thesis on unmarried mothers in this part of Sweden, this was typical of the fate of unmarried mothers at that time: If they eventually found a man to marry, with whom they had not conceived a child prior to the marriage, they were allowed to take with them at most one child to the new household.

1.1. Prologue to the present study

In a first attempt to identify the FF, the author wanted to test the possibility that the father was indeed JL (who had once claimed to be the father, but then denied it) or one of the men in F's foster family, or possibly some other man in FM's vicinity. For this purpose, the author tested both his atDNA and Y-DNA, and his strongest atDNA matches were identified. The genealogical relationship to these DNA matches were then explored by comparing the known branches of the author's family tree (i.e., his mother's branch and the known part of the FM branch) with the family trees of his atDNA matches, in a search for genealogical relationships that could explain the amount of DNA shared. As the author's FM, MF and MM and their family branches were relatively well known, it was possible to tie most of the closest DNA matches to these family branches. This was facilitated by the fact that the author's mother's family came from another part of Sweden, with a low probability of overlaps between her family tree and that of the author's father. In this way, genealogical relationships were identified for most of these DNA matches that were sufficiently close to explain the amount of DNA shared (according to data from Bettinger's [2020] Shared cM Project 4.0.].

Left, however, was a subgroup of relatively strong DNA matches where no sufficiently close genealogical relationships were found that could explain the amount of shared DNA. They were referred to as the *unplaced* DNA matches, and it was assumed that they were probably related to the author via his unknown FF.

At this stage of the research, it was still assumed that the FF was probably identical to one of the already designated candidates (i.e., either JL or one of the males in the foster family), or alternatively some man who was living or working in the FM's vicinity at the time of conception in 1912. Several such hypotheses were tested by

searching these persons' family trees for genealogical connections to the families of the unplaced DNA matches, but without success. In this way, both JL and the males in the foster family could be excluded as extremely unlikely FF candidates. The father in the foster family came from another part of Sweden, and no ancestors from his family could be found in any of the unplaced DNA matches' family trees. Although JL did have his roots closer to Västerbotten county, the same was true with regards to his family tree: no single overlap could be found between his family tree and the family trees of any of the unplaced DNA matches within a genealogical distance of six generations back in time.

The exclusion of JL as a possible FF candidate was all the more remarkable in view of the fact that the paternity trial reached an end in 1920 when FM swore on the Bible that she had been sexually involved with JL during the relevant period and that he could therefore be the father (Eckeryd, 2017; Nordmaling District Court, Fall Session 1920, § 14, 133). Although there are many possible explanations of why she did this, one plausible explanation is that she sincerely believed that JL was the father, although she had also been sexually involved with someone else during the relevant period but did not want to disclose this because she was newly married and feared for the consequences if her husband would find out.

Another approach that was tried at this early stage of the process was to search for a likely FF candidate among men who were living or working in FM's vicinity. A search for candidates were made by consulting the parish book for Nordmaling. Among a tenfold of men who appeared to be possible candidates, because they were in FM's vicinity at the time of conception in 1912, only one could be shown to be genealogically relatively closely related to *some* of the unplaced DNA matches, although not to such an extent that it made him appear as a likely FF candidate.

This lack of success led to a re-start with a new approach, where all previous suggestions about the identity of the FF were abandoned in favor of a procedure that involved an *analysis of genealogical junctions among the unplaced atDNA matches* without any presuppositions of the identity of FF. One purpose of the present paper is to describe this procedure and how it was used to identify the FF's family.

1.2. Genealogical junctions

The basic idea was that an exploration of how the unplaced DNA matches were *genealogically related to each other* would make it possible to identify *genealogical junctions*, in the form of couples (i.e., a man and a woman) probably no longer than 5 generations back in time that were common ancestors to subsets of these unplaced DNA matches. The original assumption was that these couples were likely to be ancestors to FF, and that the identification of FF would require the identification of at least two independent subsets of unplaced DNA matches defined by two different genealogical junctions, where one would represent the FF's mother's family and the other would represent the FF's father's family. If more than two genealogical junctions were identified it was expected that an analysis of the *connectedness* between these genealogical junctions (e.g., in the form of marriages between descendants) would make it possible to form a hypothesis about the identity of the FF, which could then be tested by DNA analyses. More specifically, the assumption was that the lines forward in time from the genealogical junctions would converge in a new couple that would be hypothetically identifiable as the FF's parents, and the idea was to test this hypothesis by asking living descendants of theirs to test their DNA. Partly similar methods have been used successfully in forensic contexts (e.g., Tillmar et al., 2021).

A complication, however, was that although the author's father's biological mother (FM) was known there were some gaps also in her family tree. Of special importance, her MF Carl Johan Forssén's (*1828) father was

unknown (see Figure 1). This meant that some of the unplaced atDNA matches might alternatively be explained by genealogical relationships via the unknown FMMFF. It was therefore concluded that the search for genealogical junctions in the family trees of the unplaced DNA matches had to include the identification of at least three different genealogical junctions. This might make possible also the formulation of a hypothesis about the identity of the FMMFF.

Figure 1.
The author's father's pedigree, as known at the start of the investigation[2].

TWO GENERATIONS BACK FROM THE AUTHOR
FF: Unknown
FM: Nanny Elina Ekholm *1883-11-10 Nordåker, Vännäs, Västerbotten. †1948-09-03 Robertsfors, Västerbotten

THREE GENERATIONS BACK
FFF/FFM: Unknown
FMF: Erik Ekholm *1845-12-12 Östanå, Vännäs, Västerbotten. †1920-11-13 Hörneå, Hörnefors, Västerbotten
FMM: Maria Kristina Forssén *1855-03-16 Överboda, Umeå, Västerbotten. †1915-09-16 Hörneå, Hörnefors, Västerbotten

FOUR GENERATIONS BACK
FFFF/FFFM/FMFF/FFMM: Unknown
FMFF: Johan Georg Ekholm *1821-06-13 Gumboda, Nysätra, Västerbotten. †1905-02-05 Nordåker, Vännäs, Västerbotten
FMFM: Anna Anna Elisabet Lundberg *1820-09-19 Norrmjöle, Umeå, Västerbotten. †1886-11-18 Nordåker, Vännäs, Västerbotten
FMMF: Carl Johan Forssén *1828-02-15 Björnlandsbäck, Vännäs, Västerbotten. †1865-06-08 Granlund, Vännäs, Västerbotten
FMMM: Susanna Forssén *1826-06-24 Överboda, Umeå, Västerbotten. †1897-02-01 Granlund, Vännäs, Västerbotten

FIVE GENERATIONS BACK
FFFFF/FFFFM/FFFMF/FFFMM/FFMFF/FFMFM/FFMMF/FFMMM: Unknown
FMFFF: Göran Johansson Ekholm *1789-12-31 Forsa, Hälsingland. †1859-04-14 Östanå, Vännäs, Västerbotten
FMFFM: Margareta Larsdotter *1779-01-14 Gumboda, Nysätra, Västerbotten. †1841-07-29 Östanå, Vännäs, Västerbotten
FMFMF: Fredrik Lundberg *1777-08-12 Korsholm, Vasa, Finland † 1837-10-19 Norrmjöle, Umeå, Västerbotten
FMFMF: Lisa Greta Persdotter *1779-09-14 Gubböle 2, Umeå, Västerbotten. † 1851-02-14 Norrmjöle, Umeå, Västerbotten
FMMFF: Unknown
FMMFM: Greta Stina Olofsdotter *1800-01-29 Hjuken, Vindeln, Västerbotten. †1890-09-11 Nylandsnäs, Vännäs, Västerbotten
FMMMF: Erik Eriksson Forssén *1786-02-07 Nyåker, Nordmaling, Västerbotten. †1867-12-12 Överboda, Umeå, Västerbotten
FMMMM: Charlotta Andersdotter *1790-05-22 Överboda, Umeå, Västerbotten

It was assumed that any hypothesis about the identity of FMMF's father would have to be tested in another way than the FF hypothesis. Testing the hypothesized identity of an unknown ancestor two generations back in time typically requires access to new atDNA data from his children or grandchildren. In the case of the unknown FF this would be a real possibility if children or grandchildren of a hypothetical FF could be identified and were willing to test their DNA. In the case of the FMMFF, however, no living children or grandchildren, or even great-grandchildren could realistically be expected to be found. On the other hand, the probability of finding great-great-grandchildren or great-great-great-grandchildren *who had already tested their atDNA* was judged to be relatively good. If so, it might be possible to test the hypothesis by means of *an analysis of DNA segments shared*

[2] In 19th century Sweden, family names were still uncommon and instead patronymics were in common use. This meant that both boys and girls got their surnames after their father's given name. For example, if the father's name was Olof Jonsson his sons got the surname Olofsson (i.e., Olof's son) and his daughters got the surname Olofsdotter (i.e., Olof's daughter). The latter was the case with Carl Johan Forssén's (*1828) mother, Greta Stina Olofsdotter (*1800); her father's name was Olof Jonsson, and her brothers accordingly got the surname Olofsson, whereas she bore the name Olofsdotter. So-called illegitimate children did not naturally receive any surname. The author's FMMF was given the name Carl Johan when baptized but acquired the surname Forssén first upon getting married; this was actually his wife's surname. Forssén was originally a soldier's name that her father Erik Eriksson Forssén (*1786) got while serving in the army, and that was then preserved as a family name in later generations. This was a typical origin of many family names at that time; the family name Ekholm of the FMF branch of the family was originally also a soldier's name given to the FMFFF Göran Johansson Ekholm (*1789) while serving in the army.

by descendants of FMMF Carl Johan Forssén and his hypothesized father. As a complement to searching for descendants of Carl Johan who had already tested their DNA, it was also possible to contact first and second cousins of the author who were his descendants and ask if they were willing to test their DNA.

To summarize, if a hypothetical FMMFF could be identified by the analysis of genealogical junctions, it was assumed that *segment triangulation* could be used to test this hypothesis. More specifically: If the hypothesis was correct, the author would be likely (1) to share DNA segments with DNA matches who were descendants of this hypothetical FMMFF and that (2) *also* were shared with DNA matches who were descendants of FMMF Carl Johan Forssén (*1828).

1.3. Segment triangulation

Segment triangulation is defined by Bettinger (2016) as a technique used to identify the ancestor or ancestral couple potentially responsible for a DNA segment shared by three or more descendants of that ancestor or ancestral couple. As Thomas (2021) formulates it, this is especially relevant to establishing distant relationships and "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" (p. 29). The procedure starts with the identification of an atDNA segment that the person A shares with at least two persons: B and C. A basic requirement is that all the included matches match each other (e.g., A-B, A-C, and B-C); this is essential to rule out the possibility that what looks like an apparent segment triangulation may be spurious because A shares the segment on one chromosome with B and the segment in the same location on the other chromosome in the chromosome pair with C.

The persons who share a given DNA segment are referred to as a triangulated group (TG), and when such a group has been identified the next step is to use genealogical research to identify common ancestors (CA) to the members of this TG. As pointed out by Thomas (2021), *independent lineages* are 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" (p. 71). Applied to the present research question, this means that descendants of FMMF Carl Johan Forssén (*1828) must share DNA segments to a hypothetical FMMFF via descendants of at least two of that hypothetical FMMFF's other children if the triangulation is to count as evidence of paternity.

Yet another consideration when using segment triangulation is that the strongest evidence for common ancestry is to be found when a series of *intermediate MRCAs* (most recent common ancestors) can be identified for DNA matches in each generation back in time (Bartlett, 2016). With a large TG, the ideal goal may even be described as *tracing the history of the DNA segment*, rather than as identifying one common ancestor or ancestral couple. Although this procedure involves genealogical research for the purpose of identifying MRCAs with each individual in the TG, these MRCAs will usually be found at different genealogical distances for different individuals. If the TG is sufficiently large, it may ideally be possible to identify a succession of MRCAs, generation for generation back in time, that provides a hypothetical historical reconstruction of how the DNA segment has "travelled" in time from ancient times to the present.

The important thing here is not to find matches who share strictly *identical* DNA segments, but rather who show *overlapping* of DNA segments that may be of different lengths. For example, the DNA segments shared with relatively close cousins typically tend to be larger than the segments shared with more distant cousins. The genealogical distance to the MRCAs with each individual in the TG can be expected to correlate with the size of

the DNA segments; the larger the DNA segment shared, the closer in time the MRCA may be expected to be found. No perfect correlation should be expected, however, as the transmission of DNA between generations involves a lot of randomness.

1.4. Purpose of the present study

To summarize, the purpose of the present study was

- to search for genealogical junctions (GJ) among unplaced atDNA matches
- to analyze the connectedness between these GJs in order to formulate hypotheses about the FF and FMMFF, and
- to test these hypotheses by means of further atDNA tests, Y-DNA testing, and segment triangulation.

2. MATERIAL AND METHODS

2.1. Genealogical data

There are systematic population registers in Sweden since the 16th century. In addition, a new kind of population register was introduced by a church law in 1686, whereby the country's priests were required to visit all the roots (districts) in the parish every autumn. All residents of the parish were required to participate in these house interrogations. Failure to attend could be punished with a fine and, if it was repeated, by having to sit in the church log outside the church. The purpose of the new church law was to gain better control over the population, both to ensure that people had the correct Christian doctrine and to facilitate the selection of soldiers for war service. In connection with these house interrogations, the priests were obliged to keep books of the residents of the parish, so-called parish household records.

These parish household records show variations in form but are generally organized in books, with one line for each individual and columns for the individual's name, date and place of birth, marriage, death, and migration from one parish to another. Sometimes these books also contain notes about reading or writing skills, an account of when each individual took communion, and a separate column about conduct, such as for example being convicted of a crime. These parish household records are now digitalized and are available online. The author had access to these via a subscription to *ArkivDigital* (https://www.arkivdigital.se), which is Sweden's largest and most extensive online archive of digitalized original records for genealogists.

The author, however, primarily used the full version of the genealogical database *Kråken.se*, because it provided quick access to large amounts of systematically organized genealogical information of the kind that was of primary interest in this study. This database, which was developed by the genealogist Sven-Erik Johansson, contains genealogical data on more than 600 000 persons from southern Västerbotten and northern Ångermanland (i.e., the most important geographical area for the present study). *Kråken.se* is the most complete existing genealogical database covering this part of Sweden from earliest documented time to the end of the 19th century. These data are arranged familywise with information about the time and place of birth, death, and marriage of the individual, and also contain references to the volumes and pages of the parish household books from which genealogical data were retrieved.

Pedigrees in Word format (of the form illustrated in Figure 1) were developed based on the information that the DNA matches provided themselves, in combination with information from *Kråken.se* and *ArkivDigital*. All

information was checked by consulting the digitalized archives in *ArkivDigital* before being included in the present study. Names and birth dates were entered as they were spelled and organized in *Kråken.se* to make the pedigrees easily searchable to find specific persons by using the search function in Word.

Information from US censuses, passenger lists from ships going between Sweden and the US, American muster cards from the first and second world, and other genealogical information from the US were accessed via a subscription to *Ancestry* (https://www.ancestry.se).

2.2. DNA data

The author tested his atDNA at three companies: *MyHeritage, Family Tree DNA* and *Ancestry*. On September 1st, 2023, he had 30,073 matches at *MyHeritage*, 11,282 matches at *Ancestry*, and 9,122 matches at *Family Tree DNA*. To get more complete information about the F's DNA, the author's brother was also asked to test his atDNA. Because MyHeritage produced the most matches, he tested his atDNA at *MyHeritage*; on September 1st, 2023 he had 27,713 matches. When other relatives were asked to test their atDNA this was also done on MyHeritage. As to Y-DNA it was tested by a Big-Y test at *Family Tree DNA*, both for the author and for one relative.

2.3. Procedure

The investigation proceeded in four steps: (1) the identification of unplaced DNA matches; (2) the search for genealogical junctions; (3) the generation of hypotheses, and (4) the testing of hypotheses.

2.3.1. The identification of unplaced DNA matches

In the first step, the two brothers' closest atDNA matches were identified, and the genealogical relationships to them were searched for by a comparison of the brothers' pedigree to the DNA matches' pedigrees. The closest DNA matches were defined as those who shared the most DNA with the two brothers *in combination*. This was computed as the total cM of all the DNA segments (with a size of at least 6 cM) that each DNA match shared with the two brothers. To illustrate: If a certain DNA match shared 50 cM with Brother 1 on chromosome 1 and 50 cM with Brother 2 on chromosome 2, the total shared sum would be 100 cM. But if a DNA match shared 50 cM with each brother on the same segment of chromosome 1, the total shared sum would be 50 cM.

Those of the closest DNA matches (>100 cM combined for both brothers) for which there were available family trees, but for which no explanatory genealogical relationship could be found by a comparison with the known branches of the author's family tree (i.e., his FM, MF and MM branches) within at least six generations back in time, were named *unplaced DNA matches*. To facilitate the detection of common ancestors between DNA matches, all family trees were organized in the form of Word documents in the same way as in Figure 1, to make them easy to search by entering the name and birth date of ancestors in the search field of the Word program.

2.3.2. The identification of genealogical junctions

In the second step, the family trees of the unplaced DNA matches were compared to each other, to find common ancestors no longer than five or six generations back in time. Ancestors who were common to at least two of the unplaced DNA matches were referred to as *genealogical junctions*. The task was set at finding *at least* three such genealogical junctions, the three "strongest" ones. The *strength* of genealogical junction was defined in terms of the number of unplaced DNA matches who had them in their family trees.

Here it is important to note that the strength of a genealogical junction depends also on the number of descendant branches, and that this is a potential source of error that needs to be taken account of. For example, it is quite possible that a certain couple will be identified as a genealogical junction simply because they have *a very large number of now living descendants* that have tested their DNA, despite their not being the source of the shared cM (which might instead lie one or more generations back in time).

At the other end of the spectrum, the analysis of genealogical junctions may also fail if an ancestral couple that is highly relevant to the research in question has *very few* descendants. The fewer descendants an ancestral couple has, the less likely they are to turn up as a genealogical junction in the comparison between the family trees of DNA matches, simply because they have so few descendants that can test their DNA. An ancestral couple with very few descendants is therefore less likely to be identified by an analysis of genealogical junctions. These limitations to the method are discussed further in section 4.2 of the Discussion.

2.3.3. The generation and testing of hypotheses

In the third stage the genealogical junctions that had been identified were used to generate hypotheses about the unknown ancestors. Hypotheses were generated by searching primarily for (1) convergent lines from the genealogical junctions forward in time in the form of marriages between their descendants, and if no such convergent lines could be found for (2) geographical affinities between places of residence of persons that were relevant to the investigation.

In the fourth stage the hypotheses were tested. This was done in two different ways. In the case of the author's FF the hypothesis was tested by contacting descendants of the hypothesized FF, asking them if they would like to test their atDNA and Y-DNA. In the case of the author's FMMFF the hypothesis was primarily tested by segment triangulation methods, as described below. To increase the likelihood of finding segment triangulations, two close relatives in the FM branch of the family were asked to test their atDNA.

2.3.4. Segment triangulation

A basic requirement in segment triangulation is that all the included matches match each other on the relevant DNA segment (e.g., A-B, A-C, and B-C). MyHeritage provides each tested person A with access not only to information about the segments that A shares with B and C, but also information about the segments that B and C share with each other (provided that B and C have explicitly chosen to make that information available to their DNA matches).

As implemented in the present study to test the FMMFF hypotheses, two sets of DNA matches were identified: (1) DNA matches who were descendants of Carl Johan Forssén (*1828), and (2) DNA matches who were descendants of his hypothesized father. The basic assumption was that to the extent that atDNA segments could be found that were shared *both* with descendants of Carl Johan Forssén *and* with his hypothesized father, this would count as evidence in favor of the hypothesis. The more such segments that could be found, the larger they were, and the better the conditions of independent lineages and intermediate MRCAs were satisfied (see above in section 1.3), the stronger would the evidence be.

MyHeritage provides information not only about how much DNA one shares with the DNA matches (and on which chromosomes) but also about how much DNA these matches share with each another (and on which chromosomes). To find as many descendants as possible of both the FMMF and the hypothesized FMMFF, an iterative search process was used where DNA matches who had the FMMF or the hypothesized FMMFF in their family trees were searched for their closest DNA matches in turn, comparing their family trees to see which of them also had the FMMF or the hypothesized FMMFF in their family trees. In this way, the number of descendants of both the FMMF and the hypothesized FMMFF was successfully multiplied.

The chromosome browser in MyHeritage includes detailed information about DNA segments with a size of at least 6 cM that are shared with one's atDNA matches. Information about atDNA segments of interest was exported to the program DNA Painter (https://dnapainter.com) to provide an illustrative view of how these DNA segments overlapped with each other at each chromosome. The cut-off for including DNA segments in DNA Painter was set at the size of 7 cM.

2.3.5. Ethical considerations

To preserve confidentiality, all DNA matches were given code names. These code names were constructed as a combination of a quasi-randomly generated name and the totally shared cM. To illustrate: The code name *Brian-100* would mean (1) that the person's real name was *not* Brian, and (2) that the author and his brother together shared 100 cM with him.

In addition, because the degree of anonymity of the relatives at 1st and 2nd cousin level were lower than for the other DNA matches, they were asked for their consent to refer to them under these specific code names. The manuscript was sent to them, and they were asked if they wanted to change anything to increase their degree of anonymity; all of them gave their consent to referring to them by the given code name.

When individuals test their atDNA at MyHeritage, they choose how much information they want to make available to their DNA matches. For example, they can choose if they want to make information available about the DNA segments that are shared with their matches. They also choose if they want to make genealogical information available about their family tree. For some of the author's DNA matches no family tree information was made available at MyHeritage, and for some no information about DNA segments was available; they were not included in the study.

3. RESULTS

The results are presented in five sections, of which the first four describe the results of the four different steps in the research process. In the first section, the list of the closest unplaced DNA-matches is presented. In the second section the four "strongest" genealogical junctions (GJ) are described and analysed to see if they represent independent lineages. The third section contains an exploration of the possible connectedness between these GJs in the form of marriages between their descendants and describes how this connectedness was used to generate hypotheses. The fourth section describes the testing of these hypotheses by means of atDNA, Y-DNA-testing, and segment triangulation methods. The fifth and final section contains an evaluation of the genealogical junctions method.

3.1. The closest unplaced DNA matches

A cut-off of 100 cM was chosen for how much atDNA the unplaced DNA matches had to share with the author and his brother together. In total, nineteen DNA matches who passed this cut-off were found at MyHeritage (see Table 1). The author also tested his DNA at Family Tree DNA and Ancestry, but this resulted in few strong DNA matches, and no additional match that passed the cut-off. To explore the usefulness of this cut-off and whether it might be more useful with a lower cut-off, an alternative cut-off set at 90 cM was also tested; this, however, did not contribute to any new findings of importance for the results.

Table 1 lists the nineteen unplaced DNA matches. The table also describes the amount of atDNA they shared with each brother and the two brothers in common, as well as the size of the two largest shared DNA segments, and the number of shared segments larger than 6 cM.

Table 1.
The closest unplaced DNA matches of the two brothers, and the amount of atDNA they shared in cM, the size of the two largest shared segments, and the number of shared segments, as reported by MyHeritage.

Code namn	**Brother 1 cM**	**Brother 2 cM**	**Total cM**	**Two largest shared segments (cM)**	**Number of shared segments (> 6 cM)**
Sofia-150	55	119	150	39 and 32	11
Thore-147	106	113	147	64 and 28	9
Eivor-144	89	83	144	33 and 20	11
Maja-132	52	102	132	24 and 18	11
Cesar-128	117	54	128	34 and 34	7
Willy-124	59	118	124	27 and 24	9
Igor-124	68	112	124	42 and 37	7
George-122	75	58	122	25 and 24	10
Austin-120	91	82	120	23 and 21	12
Tora-116[a]	95	74	116	16 and 14	12
Clara-114	24	114	114	40 and 18	7
Elisabeth-115	44	88	111	31 and 24	7
Axel-112	33	112	112	35 and 24	7
Marcus-111	65	79	111	39 and 16	9
Bruno-108	56	68	108	51 and 19	5
Ellie-106	59	63	106	33 and 15	9
Bianca-102	102	61	102	62 and 40	2
Ludvig-101	67	95	102	55 and 11	7
Elvira-101	40	85	101	17 and 15	10

Note. Brother 1 = the author; Brother 2 = the author's brother.
[a] *Tora-116* also tested her DNA at *Family Tree DNA*, where she was reported to share 90 cM with Brother 1.

3.2. Genealogical junctions (GJ)

The aim here was to find at least three ancestral couples no longer than five or six generations back in time who were common to at least two unplaced DNA matches (i.e., had a "strength" of at least 2), and in this sense constituted genealogical junctions (GJ). This was done by searching the Word family tree documents for persons with a specific name and birth date. Four GJs were found that had a strength of 3 or more (i.e., who had three or more of the unplaced DNA matches in their family trees). Some additional GJs were also identified that had two

unplaced DNA matches in their trees. But because four GJs had already been identified with a strength of 3, the choice was made to stay with these four. They were the following:

3.2.1. Genealogical junction 1 (GJ-1)

The strongest genealogical junction (in terms of the number of unplaced DNA matches who shared it) was a couple from southern Nordmaling born in the first decade of the 19th century: Matts Olofsson (*1806) and Maja Greta Andersdotter (*1803). Eight of the 19 unplaced DNA matches had this couple in their family tree: *Sofia-150, Maja-132, Igor-124, Tora-116, Clara-114, Elisabeth-115, Marcus-111,* and *Ludvig-101.*

Matts Olofsson (*1806) was a farmer in Ava, Nordmaling, but he was born in Bodum, a small village in Grundsunda parish, south of Nordmaling. His wife Maja Greta Andersdotter (*1803) was born in Långed, a village in the southern part of Nordmaling. After being married they moved to Ava, where they had ten children during the years 1829-1848. Of these, eight survived into adulthood and formed their own families. All in all, they had 58 grandchildren.

3.2.2. Genealogical junction 2 (GJ-2)

The next strongest genealogical junction was a couple from another part of the Nordmaling parish: Erik Olofsson (*1782) and Anna Katarina Jakobsdotter (*1788). They were found in the family trees of four of the unplaced DNA matches: *Eivor-144, George-122, Austin-120* and *Elvira-101.* Erik Olofsson (*1782) was a farmer in Mullsjö. He married a woman from the neighbouring village of Örsbäck in 1805, and they had fifteen children in Mullsjö during a 27-year period from 1806 to 1833. Ten of their children grew up to form families of their own. All in all, they had 67 grandchildren.

Interestingly, there was no overlap between the two subsets of DNA matches from GJ-1 and GJ-2 in terms of the DNA matches that had these couples in their family trees. In other words, these two subsets of DNA matches obviously represented two *independent* genealogical junctions.

3.2.3. Genealogical junction 3 (GJ-3)

The third strongest genealogical junction was a farmer couple from Ava in southern Nordmaling: Johan Petter Johansson (*1819) and Maria Karolina Mattsdotter (*1829). They had three children, of which two formed their own families, and altogether they had 25 grandchildren. This couple was found in the family trees of three of the unplaced DNA matches: *Sofia-150, Igor-124,* and *Clara-114.*

Here it may be noted that this subset of DNA matches overlapped with that from GJ-1. They actually constituted a subset of the DNA matches from GJ-1: three of the eight DNA matches from GJ-1. This meant that it did not constitute an *independent* genealogical junction. A look for genealogical connections revealed that the woman in this couple, Maria Karolina Mattsdotter (*1829), was a daughter of the couple from GJ-1.

3.2.4. Genealogical junction 4 (GJ-4)

The fourth strongest genealogical junction was a couple from Vännäs, which is about 50 kilometers to the north of Nordmaling. This couple, Jon Pehrsson (*1795 in Berg, Vännäs) and Anna Beata Vilhelmsdotter Berggren (*1791 in Vännäs) was found in the family trees of three of the unplaced DNA matches: *Cesar-128, Bruno-108*

and *Bianca-102*. If the cut-off was lowered to 90 cM one further DNA match was added. There was no overlap between this subset of DNA matches and any of the other three; it thus constituted an independent genealogical junction.

In contrast to the three previous couples, this couple was not very stationary. Jon Pehrsson (*1795) was born in the village of Berg, close to Vännäs, but in his youth he moved to the neighbouring village of Kolksele, where he worked as a servant on a farm. Eighteen years old he married one of the farmer's daughters, the four-year older Anna Beata (*1791). They later moved to Pengsjö, another small village close to Vännäs, and then to Högland, another neighbouring village. Jon Pehrsson worked as a farmer but was also involved in building the church in Vännäs in the 1820s. All in all, this couple had 13 children over a 21-year period from 1814 to 1835, ten of which grew up to have their own families, and they had 64 grandchildren. After the death of his wife, Jon Pehrsson re-married at an age of 81 years, and had 12 years with his second wife, until he died in 1889 at an age of 93.

3.3. The generation of hypotheses

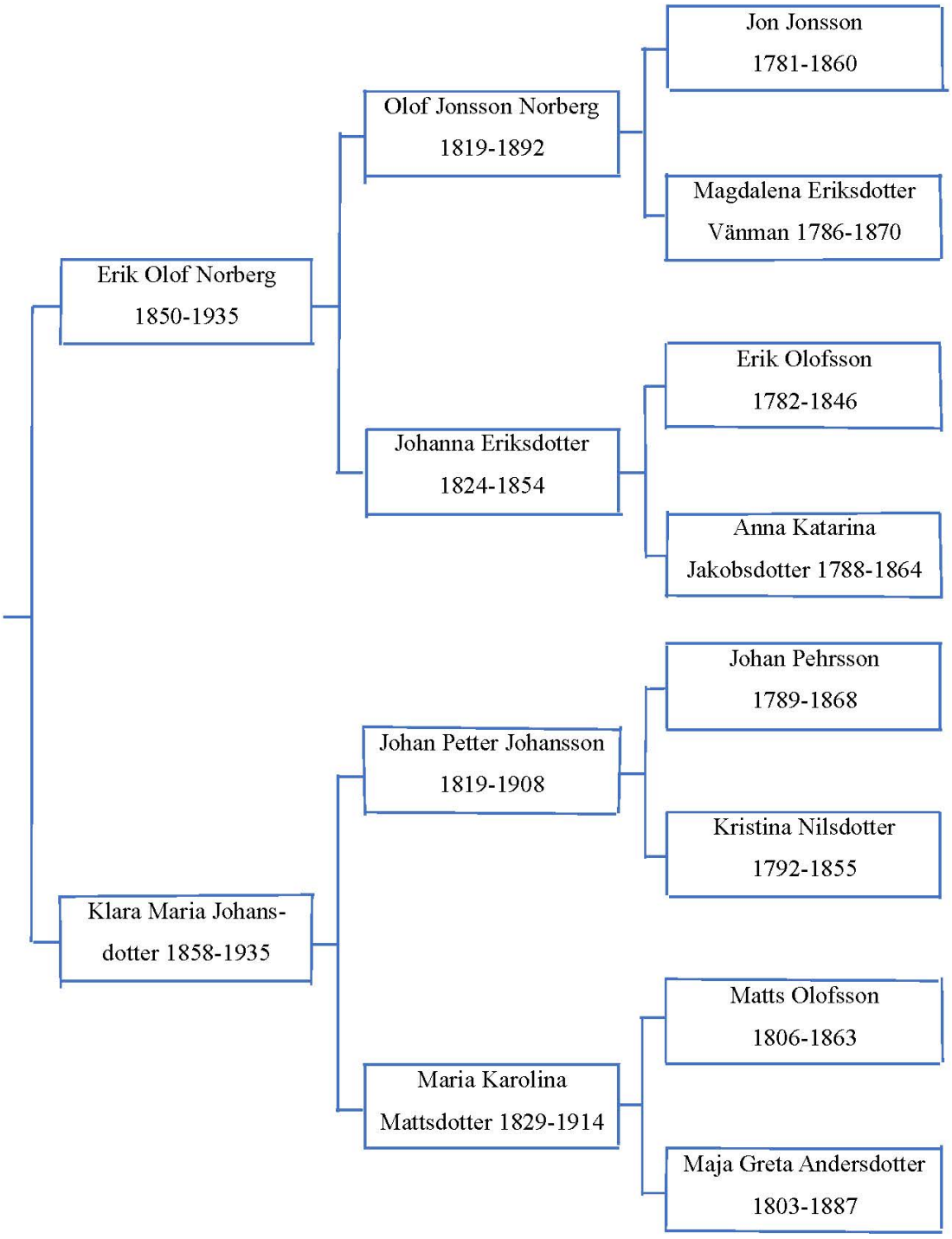
3.3.1. The FF Norberg hypothesis

To generate hypotheses about the identity of the author's FF, convergent lines from the four genealogical junctions were searched for, in the form of marriages between descendants. As already described in section 3.2.3, GJ-3 was not independent from GJ-1, as the three DNA matches whose family trees converged in GJ-3 was a subset of the eight DNA matches from GJ-1. The woman in the GJ-3 couple, Maria Karolina Mattsdotter (*1829 in Ava, Nordmaling), actually was daughter of the couple from GJ-1.

The next question was if a connection between these two hierarchically related genealogical junctions GJ-1 and GJ-3 could also be established with two other genealogical junctions GJ-2 and GJ-4, or at least one of them. A search for marriages between descendants of these GJs resulted in the discovery that a daughter of the couple in GJ-3, Klara Maria Johansdotter (*1858), on June 30th in 1878 married a grandson of the couple from GJ-2, Erik Olof Norberg (*1850). Here the family lines, in other words, converged between descendants from genealogical junctions 1, 2 and 3.

No similar convergence of family lines could be found with genealogical junction 4. It therefore seemed logical to formulate the following hypothesis about the identity of the author's biological FF: he was probably a son of Klara Maria Johansdotter (*1858) and Erik Olof Norberg (*1850), farmers in Lögdeå, Nordmaling. If this was correct the FF would have the family name of Norberg, and the hypothesis was accordingly named *the Norberg hypothesis*. The genealogical relationships between the people in genealogical junctions 1-3 are shown in Figure 2.

Figure 2.
*Erik Olof Norberg (*1850) and Klara Maria Johansdotter (*1858), their parents and grandparent, and the relationships between the genealogical junctions 1, 2 and 3. Klara Maria Johansdotter's parents represent GJ-3, whereas her maternal grandparents represent GJ-1. Erik Olof Norberg's maternal grandparents represent GJ-2.*



3.3.2. The FMMFF Jon Pehrsson hypothesis

When the author's FMMFM, Greta Stina Olofsdotter (*1800) got pregnant with her son Carl Johan (*1828; the author's FMMF) in May 1827, she worked as a maid in the small village of Pengsjö, south of Vännäs. A first conjecture, therefore, was that the father could have been a man who lived in Pengsjö at that time. This small village at that time contained five farms (numbered from 1 to 5), a soldier's croft, and a saltpetre smelter. Greta Stina worked as a maid at Pengsjö 1. Living at Pengsjö 2 were Jon Pehrsson (*1795) and Anna Beata Vilhelmsdotter Berggren (*1791), the couple in genealogical junction 4, and their children who were at that time eight in number and aged from 1 to 13 years.

This demographic information in combination with Jon Pehrsson's appearance in GJ-4 made it natural to hypothesize that Jon Pehrsson (*1795) was the author's FMMFF. First, he was living at the right place at the right time. Second, he was a common ancestor of three of the author's closest unplaced DNA matches: *Cesar-128, Bruno-108,* and *Bianca-102*. Third, some additional genealogical research revealed that, although Jon Pehrsson and his family stayed in Pengsjö only for about five years, from 1824 to 1829, he and Greta Stina Olofsdotter (*1800) might have known each other from earlier on. During the spring in 1820 Greta Stina's brother Hans Olofsson (*1798) had married one of Jon Pehrsson's first cousins, Ulrika Pehrsdotter (*1792). It seemed quite likely that Greta Stina had been present at her brother's wedding in 1820, because she and her brother seemed to be rather close as siblings (as suggested by her being one of the witnesses when her brother's first child was baptized one year after the wedding). Quite possibly, Jon Pehrsson (*1795) was also present at the same wedding, as it was one of his first cousins who got married. This indirect kinship might possibly have served as a context and excuse for Jon and Greta Stina to get acquainted during their time in Pengsjö.

3.4. Testing the FF Norberg hypothesis

To test the Norberg hypothesis, all sons in the Norberg family were identified, and a search was made for their descendants. Five such descendants (grandchildren of sons in the Norberg family) were identified and four of them were asked if they were willing to test their atDNA. All four did. The author's relationship to three of the Norberg descendants (*Viveka-509, Mattias-304*, and *Niklas-290*) went via their maternal grandfather, whereas the relationship for the fourth (*Steve-168*) went via his paternal grandfather. This meant that the hypothesis could undergo additional testing by asking *Steve-168* if he was willing to test his Y-DNA, which he was. The present section starts with a summary description of the seven brothers in the Norberg family and their known descendants, and then proceeds with a description of the results of the DNA tests.

3.4.1. The seven Norberg brothers

Erik Olof Norberg (*1850) was a farmer in Lögdeå in Nordmaling parish. He and his wife Klara Maria Johansdotter (*1858) had fifteen children during a 23-year period from 1879 to 1902. Among them were ten sons. According to the Norberg hypothesis, one of their ten sons was probably the author's FF. The youngest one was born in 1902 and since he was only eleven years when the author's father was born in 1913, he could be excluded, as could also two other sons who died as children. This left seven sons for more detailed exploration. They are referred to below as candidates 1-7.

*Candidate 1: Johan Norberg (*1879).* The oldest son, Johan Norberg (*1879) emigrated to the US in 1902 and arrived in Boston, Massachusetts on April 24th. When he applied for American citizenship in 1927, he was 48 years old and had already lived in the US for 25 years. He had worked as a miner and as a sailor/engineer on ships

going to Siberia, and he had participated in the Norwegian explorer Roald Amundsen's Maud expedition through the Northeast Passage 1918-1925. But he was still unmarried. In 1938 he married a Russian woman who was a widow with two children from an earlier marriage. She was 18 years younger, but they did not get any common children. He had no known children and therefore no descendants that could test their DNA. A study of passenger lists indicated that he had visited Sweden in 1929, but nothing indicated that he had been in Sweden in 1912 at the time when the author's father was conceived.

*Candidate 2: Eric Adolph Norberg (*1881)*. The second son, Eric Adolph (*1881) also emigrated to the US in 1902. He worked as a carpenter in Minneapolis, Minnesota and married an American woman in 1905. They had three children, born in 1907, 1910, and 1916. Only the youngest of them had children of his own, one boy and one girl. Here there were descendants that could test their DNA. But there were no indications from any passenger lists or any other documents that Eric Adolph (*1881) ever returned to Sweden.

*Candidate 3: Gustaf Fritz Norberg (*1885)*. The third son Gustaf Fritz (*1885) stayed in Sweden and took over half of the homestead in Lögdeå after his parents. He married in 1914, the year after the birth of the author's father. This meant that he was still a bachelor when the author's father was conceived, and that he was a possible FF candidate. His marriage, however, remained childless, so there were no descendants that could possibly test their DNA.

*Candidate 4: August Norberg (*1887)*. The fourth son August (*1887) emigrated to the US in April 1906 and should according to the parish book in Nordmaling have stayed there until 1921, when he returned to Lögdeå and took over half of the homestead after his parents. He settled in Lögdeå and married a woman from another family in the same village. They had one daughter, who in turn had two children, which meant that there were descendants who could test their DNA. Because he was believed to have been in US from 1906 to 1921 he seemed to be a very unlikely FF candidate. However, no documents could be found from his stay in the US, apart from his name on the passenger list of the ship Caledonia that arrived in New York on May 26th, 1906. This was in stark contrast to the wealth of information that was found for his two elder emigrating brothers Johan (*1879) and Eric Adolph (*1881): information from censuses, muster cards from the first world war, etc. The possibility could not be ruled out that he had returned to Sweden earlier than 1921.

*Candidate 5: Olov Albin Norberg (*1891)*. The fifth son, Olov Albin (*1891) married in 1930, at an age of 39 years, with a woman who had two children from previous relationships, but they had no children in common. Olov Albin stayed in Nordmaling, where he was a forest worker and ditch worker. He could not be ruled out as a FF candidate, but had no descendants who could test their DNA.

*Candidate 6: Karl Elof Norberg (*1893)*. The sixth son, Karl Elof (*1893) emigrated to the US in April 1913, four months before the birth of the author's father. He stayed in the US until 1933, when he returned to Nordmaling. In this case, just as for his two oldest brothers Johan (*1879) and Eric Adolph (*1881), it was easy to find documents referring to his life in the US: registration cards from the First World War, censuses, etc. He lived in Hennepin, Minnesota, where he worked as a clerk on a firm called Sash & Door. He never married and had no descendants that could test their DNA, but he could not be ruled out as an FF candidate.

*Candidate 7: Axel Nikanor Norberg (*1896)*. The seventh son, Axel Nikanor (*1896) stayed in Lögdeå, Nordmaling and married a teacher from the neighbouring village Mo, where they settled. They had one daughter, and she in turn married and had a son, so here there was one descendant who could test his DNA.

3.4.2. Testing atDNA

Apparently, only three of the seven brothers had descendants that could test their atDNA: Eric Adolph (*1881), August (*1887) and Axel Nikanor (*1896). The author approached the two known descendants of August (*1887), the only known descendant of Axel Nikanor (*1896), and one descendant of Eric Adolph (*1881), and asked if they were willing to test their DNA, and all of them did. The results from their DNA testing, in terms of the amount of DNA shared with the author and his brother, and between themselves, is shown in Table 2.

Table 2.
The amount of DNA(in cM) shared between brothers 1 and 2 and descendants of the Norberg brothers.

	Brother 1	Brother 2	*Viveka-509*	*Mattias-309*	*Niklas-290*	*Steve-168*
Brother 1	-					
Brother 2	2728 cM	-				
Viveka-509	286 cM	372 cM	-			
Mattias-304	167 cM	208 cM	2761 cM	-		
Niklas-290	204 cM	113 cM	214 cM	203 cM	-	
Steve-168	101 cM	102 cM	162 cM	98 cM	146 cM	-

Note. *Viveka-509* and *Mattias-304* are siblings and descendants of August Norberg (*1887). *Niklas-290* is descendant of Axel Nikanor Norberg (*1896). *Steve-168* is descendant of Eric Adolph Norberg (*1881).

As seen in Table 2, *Viveka-509* was the one who shared the most atDNA with her second cousins and with brothers 1 and 2: At an average she shared 188 cM with her 2nd cousins *Niklas-290* and *Steve-168*, and 329 cM with brothers 1 and 2. Her brother *Mattias-304* shared at an average 151 cM with his 2nd cousins *Niklas-290* and *Steve-168* and 188 cM with brothers 1 and 2. That is, both grand-children of August Norberg shared even more atDNA with brothers 1 and 2 than they did with their 2nd cousins *Niklas-290* and *Steve-168.* This clearly supported the Norberg hypothesis.

As to *Niklas-290* he shared at an average 188 cM with his three 2nd cousins *Viveka-509, Mattias-304*, and *Steve-168*. He shared at an average a little less than so, 159 cM, with brothers 1 and 2. *Steve-168* was the one who shared the least atDNA with the others: at an average 135 cM with his three 2nd cousins *Viveka-509, Mattias-304*, and *Niklas-290*, and a little less than so with the two brothers: 101,5 cM. Interestingly, the amount of DNA that brothers 1 and 2 shared with their four hypothesized 2nd cousins fell well within the range of shared DNA among the Norberg descendants: an average of 190 cM for Brother 1, and 199 cM for Brother 2. This was clearly consistent with the Norberg hypothesis: both brothers shared as much DNA as could be expected to be shared with the Norberg descendants if they were their 2nd cousins.

A crucial question was if any of these three Norberg brothers could be identical with the author's FF. If so, grandchildren of that brother would be the author's *half first cousin*, whereas grandchildren of the other brothers would be the author's *second cousins*. A correct interpretation of these results requires empirical data on how much DNA one tends to share with one's half first cousins versus one's second cousins. Data on this was taken from Bettinger's (2020) Shared cM Project 4.0 and are shown in Table 3.

Table 3.
Amount of DNA shared with half first cousins and second cousins, according to Bettinger's (2020) Shared cM Project 4.0.

Genealogical relationship	**Average shared DNA**	**Range (low to high; 99th percentile)**
Half first cousins	449	156-979
Second cousins	229	41-592

First, these data indicate that neither Axel Nikanor Norberg (*1896) nor Eric Adolph Norberg (*1881) were likely FF candidates. The reason for this is that Brother 2 shared only 113 cM with *Niklas-290*, which was below the range for half first cousins (i.e., less than 156 cM), and that both Brother 1 and 2 shared even less with *Steve-168.*

Second, although it could not be conclusively ruled out that August Norberg (*1887) was the FF, this did not seem very likely. The amount of DNA that the two brothers shared with August's grandchildren *Viveka-509* and *Mattias-304* was clearly within the range for half first cousins (156-979 cM), but it was also clearly within the range for second cousins (41-592). What primarily spoke against August Norberg being the FF, however, was the results for *Mattias-304*; the amount of DNA that he shared with both Brother 1 and 2 (167 cM and 208 cM, respectively) was lower even than the average for second cousins (229 cM).

3.4.3. Y-DNA

Of the four Norberg descendants, the author was related to three via their maternal grandfathers and to one, *Steve-168*, via his paternal grandfather. *Steve-168* was therefore asked if he would be willing to test his Y-DNA. He did this by taking a Big-Y test at Family Tree DNA, and the results were classified as an "exact match". *Steve-168* belonged to the same haplogroup as the author, R-YP4123. The comparison of STRs showed that only 1 of 653 differed, and the matching of private SNPs showed only two non-matching variants. This meant that the Norberg hypothesis was corroborated also by the testing of Y-DNA.

To conclude, this first part of the analysis clearly indicated that one of the Norberg brothers was the author's biological FF, but it was not possible to decide which of them. The most likely candidates were Gustaf Fritz (*1885), Olov Albin (*1891) and Karl Elof (*1893), but none of them had any known children and therefore no descendants that could test their DNA. Although August (*1887) could not be excluded, he did not seem to be a very likely candidate.

3.4.4. Segment triangulation

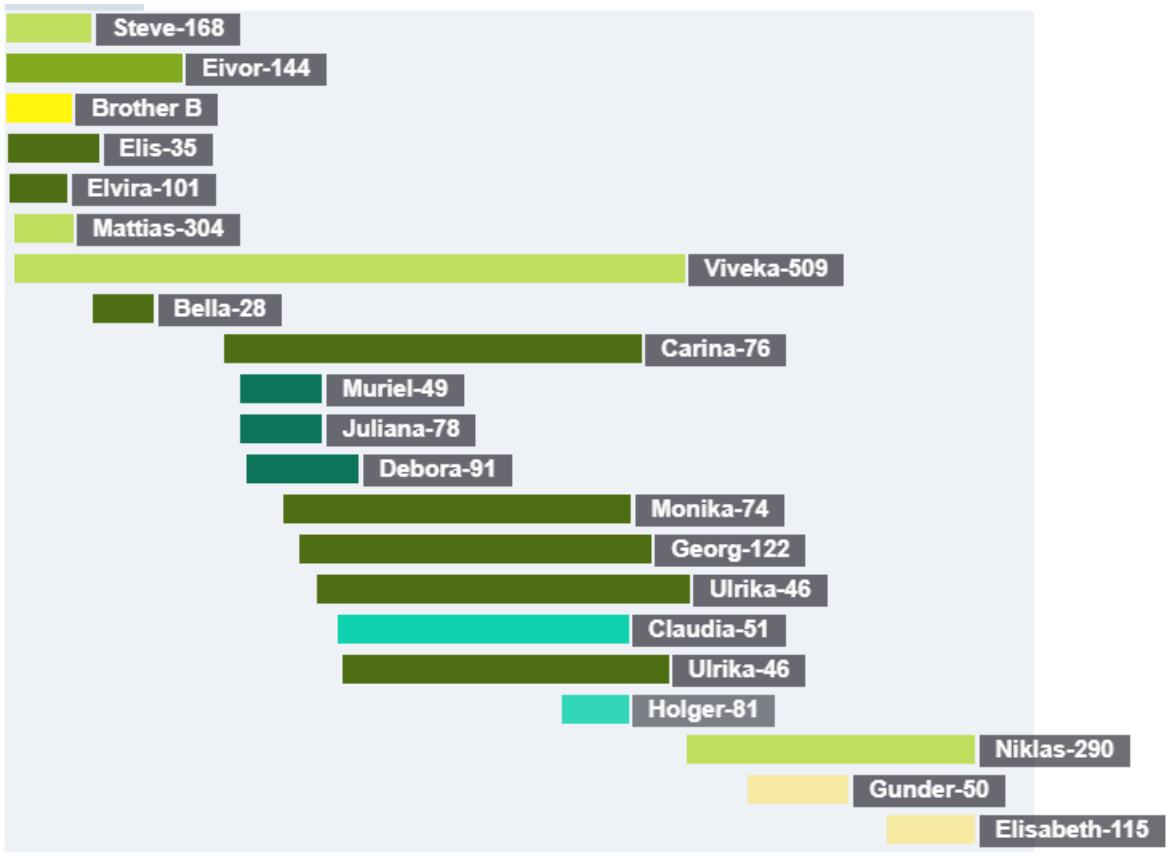
Finally, the Norberg hypothesis was tested also by means of segment triangulation. The reasoning was as follows: If the Norberg descendants *Viveka-509, Mattias-304, Niklas-290* and *Steve-168* were 2nd cousins to brothers 1 and 2, this should (1) result in the sharing of some relatively large DNA segments between the Norberg descendants and the two brothers, which (2) should also be shared (at least in part) by other DNA matches who had either Erik Olof Norberg's (*1850) or Klara Maria Johansdotter's (1858) parents or grandparents in their family trees (cf. the pedigree in Figure 2).

The testing of the hypothesis was made in the following steps: (1) The chromosome browser in MyHeritage was used to identify eleven relatively large (>30 cM) DNA segments where Brother 1 and/or Brother 2 shared DNA with at least one of the Norberg descendants. (2) The largest of these DNA segments was selected for more detailed analysis. This was a segment found on chromosome 16, where Brother 1 shared 78 cM with *Viveka-509*. (3) The chromosome browser in MyHeritage was used to identify DNA matches who shared this DNA segment (or part of it) and had ancestors to the Norberg brothers in their family trees. (4) This information was exported into DNA Painter to get an illustrative picture of the way these individuals shared DNA on chromosome 16 (see Figure 3).

As can be seen in Figure 3, of the Norberg descendants *Viveka-509*'s brother *Mattias-304* also shared 11 cM of this segment, and their 2nd cousin *Steve-168* shared 18 cM of it, both at the very beginning of the segment. Brother

2 also shared 12 cM of the segment at its very beginning. The only one among the Norberg descendants who did not share any part of this segment was *Niklas-290*; on the other hand, as seen in Figure 3, *Niklas-290* shared a relatively large segment with the author that started right where the larger segment ended and stretched almost to the end of the chromosome.

Figure 3.
*DNA segments on chromosome 16 shared between Brother 1 and descendants of Erik Olof Norberg (*1850) and Klara Maria Johansdotter (1858), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage.*



MRCA: FFF Erik Olof Norberg (*1850) and FFM Klara Maria Johansdotter (*1858)
MRCA: FFFF Olof Norberg Jonsson (*1819) and FFFM Johanna Eriksdotter (*1824)
MRCA: FFFMF Erik Olofsson (*1782) and FFFMM Anna Katarina Jakobsdotter (*1788)
MRCA: FFFMFF Olof Isaksson (*1746) and FFFMFM Margareta Olofsdotter (*1755)
MRCA: FFFMMF Jakob Jakobsson (*1760) and FFFMMM Brita Persdotter (*1765)
MRCA: FFMMF Matts Olofsson (*1806) and FFMMM Maja Greta Andersdotter (*1803)

As depicted in Figure 3, Brother 1 shared almost this entire chromosome with Norberg descendants, which indicates that he had received almost the entire chromosome from his FF. In contrast, Brother 2 shared only a small part with Brother 1 on this chromosome (12 cM at the very beginning). Brother 2, in fact, had inherited almost all DNA on his corresponding chromosome 16 from the FM branch of the family (see below in section 3.5.2).

If the Norberg hypothesis were correct, the origin of this large DNA segment on chromosome 16 should be possible to trace to ancestors of the father and/or the mother in the Norberg family. As seen in Figure 3, the results were clearly in line with the hypothesis. The large segment shared with *Viveka-509* seemed to have come from the Norberg brothers' paternal grandmother, Johanna Eriksdotter (*1824), who had received part of it from her father and part of it from her mother. This is indicated by the following findings:

1. First, a relatively large part of this segment, 33 cM, was shared with the author's 3rd cousin *Eivor-144.* Because her MRCA were Erik Olof Norberg's (*1850) parents Olof Norberg Jonsson (*1819) and Johanna Eriksdotter (*1824), this suggested that at least part of the segment originated from this ancestral couple.

2. Second, among the author's 4th cousins that have already been mentioned previously in this paper (because they belonged to the author's closest unplaced DNA matches), 23 cM of the segment was shared by *Georg-122,* and another 10 cM was shared by *Elvira-101*. Their MRCA were Erik Olof Norberg's (*1850) maternal grandparents Erik Olofsson (*1782) and Anna Katarina Jakobsdotter (*1788). This indicated that the segment, or at least parts of it, came from this ancestral couple. This hypothesis was further strengthened by the fact that several other 4th cousins with the same MRCA also shared parts of this DNA-segment: *Carina-76* (32 cM), *Monika-74* (23 cM), *Ulrika-46* (26 cM), *Elis-35* (19 cM), and *Bella-28* (12 cM).

3. Five additional DNA matches were found who had the *parents* of Erik Olofsson *1782 or Anna Katarina Jakobsdotter *1788 in their family trees. Three of these (*Deborah-91, Juliana-78* and *Muriel-49*) had Erik Olofsson's (*1782) parents Olof Isaksson (*1746 in Mullsjö, Nordmaling) and Margareta Olofsdotter (*1755 in Mullsjö, Nordmaling) among their ancestors, whereas two others (*Holger-81* and *Claudia-51*) had his wife Anna Katarina Jakobsdotter's (*1788) parents Jakob Jakobsson (*1760 in Örsbäck, Nordmaling) and Brita Persdotter (*1765 in Ängersjö, Nordmaling) in their trees. This suggested that different parts of the segment might have their origin in these two ancestral couples.

Finally, as a contrast, it can be seen in Figure 3 that the segment on chromosome 16 that was shared with *Niklas-290* (as seen in the right part of the figure) was shared also with two other DNA matches (*Elisabeth-115* and *Gunder-50*) whose family trees contained another ancestral couple: the Norberg brothers' mother's maternal grandparents Matts Olofsson (*1806) and Maja Greta Andersdotter (*1803). This suggested that the segment that was shared by *Niklas-290* probably had its origin in that branch of the family.

To summarize: the Norberg hypothesis had now been tested in three different ways, and all results were clearly in line with the hypothesis. Everything pointed to one of the Norberg brothers as being the FF.

3.5. Testing the FMMFF Jon Pehrsson hypothesis

The Jon Pehrsson hypothesis was more difficult to test, partly because the searched-for ancestor was five generations back in time, whereas the Norberg hypothesis was about an ancestor two generations back in time. In this case the hypothesis was tested primarily by segment triangulation methods, although the process was facilitated by contacting relatives in the FM branch and asking them if they were willing to test their atDNA. Two such relatives were addressed, and both were willing to collaborate: a half first-cousin one step removed (*Valter-261*) and a second cousin (*Ramona-420*). This added more DNA matches who were descendants of Carl Johan Forssén (*1828) for segment triangulation purposes.

The chromosome browser in MyHeritage presents detailed information about the overlapping of one's DNA with that of the DNA matches at each chromosome. This made it possible to explore to what extent the DNA segments that Brother 1 and/or Brother 2 shared with descendants of Carl Johan Forssén (*1828) were also shared by descendants of his hypothesized father Jon Pehrsson (*1795). If such overlapping between DNA segments were found this would count as support for the hypothesis.

To test the hypothesis, a search was first made for (1) all DNA matches who were descendants of Carl Johan Forssén (*1828), and (2) as many DNA matches as possible who were descendants of Jon Pehrsson (*1795). (See also section 2.3.4 under Methods and materials.) Table 4 lists the DNA matches who were descendants of Carl Johan Forssén, how much DNA was shared with them, the nature of the author's genealogical relationship to them, and their lineage to Carl Johan Forssén (i.e., which of his children they descended from). As seen in the table, the matches included descendants of five of Carl Johan Forssén's children: Katarina Charlotta (*1851), Maria Kristina (*1855), Susanna Sofia (*1857), Karl (*1860), and Johan Petter (*1862).

Table 4.
*DNA matches who were descendants of FMMF Carl Johan Forssén (*1828), the amount of DNA shared with them (in cM, as reported by MyHeritage) and the nature of their genealogical relationship to the FMMF.*

Code name	Brother 1 cM	Brother 2 cM	Total cM	Genealogical relationship	Descendant of
Ramona-420	339[a]	254	420	Second cousin	FFMF's daughter Maria Kristina *1855
Valter-261	149	187	261	Half first cousin one step removed	FFMF's daughter Maria Kristina *1855
Robin-219	181	190	219	Second cousin one step removed	FFMF's daughter Maria Kristina *1855
Henrik-175	68	130	175	Third cousin one step removed	FFMF's daughter Susanna Sofia *1857
Amanda-147	99	77	147	Third cousin	FFMF's daughter Susanna Sofia *1857
Ella-129	98	129	129	Third cousin one step removed	FFMF's son Karl *1860
Abigail-125	24	101	126	Third cousin two steps removed	FFMF's son Karl *1860
Ingeborg-119	112	45	119	Third cousin	FFMF's daughter Katarina Charlotta *1851
Erik-101	48	84	105	Third cousin	FFMF's daughter Susanna Sofia *1857
Elof-99	49	50	99	Third cousin	FFMF's son Johan Petter *1862
Beata-70	42	64	70	Third cousin	FFMF's son Johan Petter *1862
Jesper-56	33	43	56	Third cousin one step removed	FFMF's daughter Susanna Sofia *1857
Kerstin-53	45	18	53	Third cousin two steps removed	FFMF's daughter Katarina Charlotta *1851
Arthur-50	35	15	50	Third cousin one step removed	FFMF's son Johan Petter *1862
Cilla-46	46	0	46	Third cousin two steps removed	FFMF's daughter Katarina Charlotta *1851
Isa-24	24	0	24	Third cousin two steps removed	FFMF's son Johan Petter *1862

[a] *Ramona-420* also tested her DNA at Ancestry, where she was reported to share 324 cM with Brother 1.
Note. Brother 1 = the author; Brother 2 = the author's brother.

Table 5 similarly shows a list of DNA matches who were descendants of Jon Pehrsson (*1795), how much DNA was shared with them, and how these DNA matches were genealogically related to Jon Pehrsson (i.e., which of his children they descended from).

Table 5.

*DNA matches who were descendants of Jon Pehrsson's (*1795) children (their names depicted in different colours) in his marriage with Anna Beata Vilhelmsdotter Berggren (*1791), and the amount of DNA shared with these matches (in cM) as reported by MyHeritage.*

Code name	Brother A	Brother B	Total cM	Jon Pehrsson's genealogical relationship to the DNA matches and name of the children they were descendants of
Cesar-128	117	54	128	FFFFF via Erik (*1820)
Bruno-108	56	68	108	MMFF via Johan (*1835)
Bianca-102	102	61	102	FMFFFF via Johan (*1835)
Jessica-92	74	34	92	FFFFF via Johan (*1835)
Kasper-83	9	74	93	FMFFF via Erik (*1820)
Oscar-80	48	40	80	FMFFF via Erik (*1820)
Elsy-78	0	78	78	FFMFFF via Erik (*1820)
Anne-74	74	--[a]	74	MFFFFF via Erik (*1820)
Jerry-69	0	69	69	FFMFFF via Erik (*1820)
Malin-69	30	39	69	MFFMFF via Vilhelm Petter (*1814) and MFMFFF via Johan (*1835)
Jörgen-68	0	68	68	MFMMF via Sofia Helena (*1834)
Erling-65	0	65	65	MFFFFF via Erik (*1820)
Liv-64	27	37	64	FFFF via Johan (*1835) and MFMFF via Vilhelm Petter (*1814)
Eileen-64	38	43	65	FFFFF via Erik (*1820)
Benjamin-62	62	0	62	FFFF via Erik (*1820)
Timmy-61	39	41	61	FMFMFF via Vilhelm Petter (*1814)
Anny-60	33	48	60	FMMMFF via Johan (*1835)
Martin-60	0	60	60	MFFFFF via Erik (*1820)
Derek-59	36	39	59	FFFMFF/ MFFMFF via Vilhelm Petter (*1814)
Hellen-56	48	39	56	FMFMMF via Sofia Helena (*1834)
Lucy-55	47	32	55	MFFMFF via Vilhelm Petter (*1814)
Blenda-49	49	12	49	MFFFMF via Maria Kristina (*1818)
Hedda-45	0	45	45	FMMFF via Vilhelm Petter (*1814)
Ragnhild-43	16	27	43	FMFMMF via Maria Kristina (*1818)
Kinna-42	42	0	42	MFMFMF via Cajsa Lisa (*1828)
Rasmus-41	0	41	41	FFFFMFF via Vilhelm Petter (*1814)
Kicki-40	40	0	40	MMMFMF via Cajsa Lisa (*1828)
Pamela-40	40	0	40	MMFMF via Cajsa Lisa (*1828)
Bertram-39	30	22	37	FMMMF via Barbro Magdalena (*1823)
Hildegard-39	39	25	39	MFFMFF via Vilhelm Petter (*1814)
Bernarda-38	37	22	37	FFFMMMF via Barbro Magdalena (*1823)
Tullia-38	38	0	38	FFMMFF/ MMFMFF via Erik (*1820)
Vicky-38	24	38	38	MFMFFFF via Erik (*1820)
Li-37	0	37	37	MMFFMFF via Vilhelm Petter (*1814) and MMFMFFF via Johan (*1835)
Mats-37	37	22	37	FMMMF via Barbro Magdalena (*1823)
Joe-33	18	33	33	MMMFF via Johan (*1835)
Rodney-32	29	26	32	MFMFF via Barbro Magdalena (*1823)
Quinn-29	29	0	29	MMFMF via Cajsa Lisa (*1828)
Marian-28	28	0	28	MFMFMF via Cajsa Lisa (*1828)
Sofie-28	28	0	28	FMMMF via Barbro Magdalena (*1823)
Andres-27	14	29	27	MFFFFF via Johan (*1835)
Joanna-23	23	0	23	MFFFFF via Johan (*1835)
Gregor-20	20	0	20	FMFFF via Erik (*1820)
Katrin-20	20	0	20	MFMFFF via Erik (*1820)
Lukas-19	19	0	19	FFMMMF via Anna Margareta (*1821)
Malena-19	0	19	19	MMFMFF via Erik (*1820)
Sture-18	18	0	18	FMFFF via Erik (*1820)
Matti-17	17	0	17	FMMFF via Johan (*1835)
Bernie-16	16	0	16	MFMFMMF via Maria Kristina (*1818)

[a] *Anne-74* had only tested her DNA at Family Tree DNA, and Brother 2 had not tested his DNA there.

Note. Brother 1 = the author; Brother 2 = the author's brother.

As seen in Table 5, the matches included descendants of eight of Jon Pehrsson's children: Vilhelm Petter (*1814), Maria Kristina (*1818), Erik (*1820), Cajsa Lisa (*1820), Anna Margareta (*1821), Barbro Magdalena (*1823), Sofia Helena (*1834), and Johan (*1835). The names of the children are depicted in different colours; this information is important to establish independent lineages (Thomas, 2021) in the segment triangulations described below. More detailed information about the lineage from each of these DNA matches to Jon Pehrsson is also found in the table; as seen on the first row, for example, Jon Pehrsson was *Cesar-128's* FFFFF.

Common to all the DNA matches in Table 4 are that they are descendants of Carl Johan Forssén (*1828; the author's FMMF), and common to all DNA matches in Table 5 are that they are descendants of Jon Pehrsson (*1795) children in his marriage with Anna Beata Vilhelmsdotter Berggren (*1791). If Jon Pehrsson (*1795) was the father of Carl Johan Forssén (*1828), they would be expected to share a substantial amount of DNA, and a substantial number of atDNA segments. Accordingly, their respective descendants would also be expected to share some of these DNA segments with each other. By implication, the author and his brother (as being descendants of Carl Johan Forssén) would also be expected to share some of the segments that these two sets of DNA matches share with each other. It should be noted that this procedure requires a segment triangulation involving *three* different sets of individuals: (1) Brother 1 and/or 2; (2) other descendants of Carl Johan Forssén; and (3) descendants of Jon Pehrsson's other children.

The hypothesis was tested in the following way: First, information about DNA segments (>7 cM) shared with the DNA matches in Table 4 (descendants of Carl Johan Forssén) and the DNA matches in Table 5 (descendants of Jon Pehrsson) that was available in the chromosome browser at MyHeritage was exported to the program DNA Painter (https://dnapainter.com). This provided an illustrative view of how these DNA segments overlapped with each other at each chromosome. Second, all chromosomes were searched for segments where Brother 1 and/or Brother 2 had an overlap of DNA *both* with descendants of Carl Johan Forssén *and* with descendants of Jon Pehrsson's children. Third, to make sure that this overlapping was on the same chromosome (and not on the other chromosome in the pair) it was ensured that the matches who shared segments with brothers 1 and/or 2 also shared these segments (or at least part of these segments) with each other.

*3.5.1. Overview of the segment triangulations between descendants of Carl Johan Forssén (*1828) and descendants of Jon Pehrsson (*1795)*

Table 6 shows atDNA segments that were shared *both* with descendants of Carl Johan Forssén (*1828) *and* with descendants of Jon Pehrsson's (*1795) children. Only triangulations that involved segments where at least one DNA match shared more than 15 cM with Brother 1 or Brother 2 were included. (For example, a triangulation on segment 2 with two descendants of Carl Johan Forssén and two descendants of Jon Pehrsson was excluded, as all four segments measured only between 7 and 13 cM). The names of the descendants of Jon Pehrsson are written in different colours, depending on which children they descendent from (cf. Table 5); this makes it easier to see to what extent independent lineages were involved in the various triangulations.

Table 6.

*Segment triangulations between Brother 1 and/or Brother 2 and DNA matches who were descendants of Carl Johan Forssén (*1828) and DNA matches who were descendants of Jon Pehrsson (*1795) children. The chromosome number, segment size (cM) and start and end location for each segment is included. The different colours indicate which of Jon Pehrsson's children they were descendants of:* Vilhelm Petter (*1814), Maria Kristina (*1818), Erik (*1820), Anna Margareta (*1821), Barbro Magdalena (*1823), Cajsa Lisa (*1828), Sofia Helena (*1834),or Johan (*1835).

Chromosome No. (Brother 1 and/or 2)	Descendants of Carl Johan Forssén (*1828); segment size (cM); start and end location	Descendants of Jon Pehrsson' (*1795) children
1 (A+B)	*Ramona-420:* 41 cM [107,825,866-162,252,097] *Robin-219*: 7 cM [162,357,707-167,403,625]	*Hannes-39:* 39 cM [88,794,011-146,672,906] *Jessica-92*: 16 cM [88,794,011 - 107,823,943] *Bianca-102:* 62 cM [100,864,133-173,376,184] *Lucy-55:* 12 cM [103,756,707-113,288,057] +8 cM [154,814,917-160,678,721 *Hildegard-39*: 27 cM [154,814,917 - 176,446,394]] *Anne-74:* 8 cM [161,919,011-167,103,268] *Derek-59:* 13 cM [165,125,851 - 177,342,651]
2 (B)	*Henrik-175:* 21 cM [169,010,659-192,417,747] *Valter-261:* 14 cM [169,427,769-180,693,147] *Erik-105:* 20 cM [175,333,632-202,897,049] *Beata-70:* 21 cM [193,472,864-217,414,502] *Arthur-50:* 15 cM [196,051,485-212,888,988]	*Malin-69:* 25 cM [171,075,730-199,419,373] *Timmy-61: 26 cM:* [171,358,785 - 202,897,049] *Vicky-38:* 7 cM [191,800,905-200,930,815]
3 (A)	*Ramona-420:* 32 cM [148,317,504 - 181,443,42] + 12 cM [188,430,230 - 193,962,67]	*Bruno-108:* 17 cM [180,633,485 - 189,583,705] *Cesar-128*: 17 cM [181,692,310 - 189,948,515] *Anne-74*: 15 cM [182,372,828 - 189,590,825] *Elmar-38*: 12 cM [185,447,038 - 189,948,515]
4 (A)	*Amanda-147:* 28 cM [5,794,904-22,396,295] *Arthur-50:* 9 cM [5,794,904-8,031,590] *Ramona-420:* 15 cM [8;774,285-21,183,041]	*Kinna-42:* 27 cM [5,920,389-21,816,328]
6 (A)	*Ramona-420:* 41 cM [123,805,429-155,763,508] *Henrik-175:* 23 cM [123,805,429-143,702,113]	*Cesar-128:* 9 cM [124,975,861-133,614,576] *Derek-59:* 9 cM [124,975,861-134,090,307] *Anny-60:* 8 cM [125,219,805-133,077,063]
7 (A+B)	*Valter-261:* 22 cM [43,748 - 11,597,743]	*Bernarda-38*: 12 cM [43,748 - 6,495,017] *Verna-17*: 10 cM [915,678 - 6,026,607] *Mats-37*: 11 cM [915,678 - 6,933,726]
9 (B)	*Elof-99:* 29 cM [114,636,015 - 134,717,69] *Ramona-420*: 11 cM [125,075,782 - 133,351,106]	*Malin-69:* 15 cM [99,448,564 - 112,271,779] *Liv-64*: 18 cM [100,270,846 - 115,103,266] *Eileen-64*: 18 cM [100,507,445 - 115,430,157] *Vicky-38*: 7 cM [110,383,797 - 115,430,157]
10 (A)	*Robin-219:* 24 cM [83,286,025 - 108,878,616]	*Bernie-16*: 16 cM [83,739,866 - 99,268,810] *Ragnhild-43*: 16 cM [83,790,207 - 99,551,326] *Timmy-61*: 22 cM [90,101,723 - 114,197,642]
12 (A)	*Ramona-420:* 17 cM [52,374,207-68,138,077]	*Benjamin-62:* 55 cM [5,383,579-54,593,124] *Bianca-102:* 40 cM [23732165- 67,920,566] *Jessica-92:* 38 cM [23,732,165-67,066,115] *Sture-18*: 18 cM [29,762,556 - 53,175,287] *Lukas-19:* 19 cM [42,068,852-62,103,023]
13 (A)	*Robin-219:* 14 cM [38,363,047-48,376,814] *Ingeborg-119:* 7 cM [44,249,607 - 49,798,391]	*Jessica-92:* 19 cM [38,363,047-57,054,331] *Bruno-108:* 19 cM [38,551,272-57,054,331] *Marian-28:* 14 cM [38,551,272-49,434,635]

		Mulle-15: 15 cM [38,363,047 - 49,798,391] *Elmar-38:* 12 cM [46,287,707-61,160,649] *Eberhard-39*: 10 cM [46,655,501 - 60,218,409]
15 (B)	*Robin-219:* 26 cM [85,939,032 - 96,900,999]	*Hellen-56*: 25 cM [8,543,639 - 93,885,108]
16 (B)	*Henrik-175:* 55 cM [5,488,091-53,137,282] *Ramona-420:* 30 cM [11,930,337-34,786,294] *Valter-261:* 27 cM [12,430,207- 27,887,780] + 39 cM [77,447,299-90,233,487] *Abigail-125:* 81 cM [20,419,699-85,659,155] *Amanda-147:* 30 cM [46,538,105- 71,961,906] *Elof-99:* 14 cM [78,197,287-82,866,767]	*Elsy-78:* 14 cM [4,171,701-9,173,143] + 51 cM [17,338,502-69,892,430] *Martin-60:* 13 cM [4,388,655- 9,173,143] + 46 cM [17,338,502- 63,545,556] *Jerry-69:* 14 cM [4,388,655-9,551,490] + 55 cM [17,338,502-73,752,262] *Kasper-83:* 14 cM [4,388,655-9,366,359] + 21 cM [17,950,096-34,786,294] + 33 cM [46,538,105-73,752,262] *Malena-19:* 11 cM [5,294,643-9,173,143] + 8 cM [56,737,620-62,268,125] *Bruno-108:* 51 cM [9,184,894-56,594,499] *Hedda-45:* 9 cM [6,082,553- 9,173,143] + 21 cM [58,867,991-78,883,118] *Rasmus-41:* 10 cM [6,082,553-9,366,359] +18 cM [59,888,170-78,041,865] *Erling-65:* 9 cM [6,164,478-9,173,143] + 44 cM [21,288,101-67,409,180]
17 (A)	*Amanda-147:* 29 cM [13,222,160-38,447,569]	*Oscar-80:* 20 cM [17,570,827-38,447,569] *Anne-74:* 20 cM [18,566,664-38,955,833] *Liv-64:* 19 cM [18,626,014-38,447,569] *Cesar-128:* 19 cM [18,626,014-38,447,569]
18 (B)	*Ramona-420:* 21 cM [0,582,838-7,339,906] *Robin-219:* 13 cM [0,582,838-5,069,982]	*Jörgen-68*: 11 cM [3,042,297-6,260,004]
19 (A)	*Erik-105:* 20 cM [48,760,229-55,069,964] *Beata-70:* 17 cM [51,595,182-56,596,815] *Isa-24:* 18 cM [51,595,182-56,762,676]	*Blenda-49*: 12 cM [49,171,593-53,577,419]
20 (B)	*Abigail-125:* 21 cM [31,859,221 - 48,913,503]	*Jörgen-68*: 37 cM [14,559,429-47,778,602]
22 (A)	*Ingeborg-119* 27 cM [22,235,510 - 37,033,264] *Kerstin-53:* 12 cM [22,910,416-27,647,065] *Cilla-46:* 11 cM [23,262,118-27,647,065]	*Matti-17*: 17 cM [21,363,306-28,074,256] *Bruno-108:* 14 cM [23,262,118-29,390,049] *Jessica-92*: 17 cM [27,647,289 - 38,164,106] *Katrin-20*: 20 cM [36,241,136 - 45,772,802] *Gregor-20*: 20 cM [36,241,136 - 45,772,802] *Oscar-80*: 21 cM: [37,966,060 - 49,067,699]

Note. Brother 1 = the author; Brother 2 = the author's brother.

As seen in Table 6, segment triangulation with descendants of Carl Johan Forssén (*1828) *and* descendants of Jon Pehrsson's (*1795) children was found on 17 chromosomes. Six of these (on chromosomes 4, 7, 15, 18, 19, and 20), however, failed to include independent lineages as each of them included only one of Jon Pehrsson's children. Six others did include independent lineages with two of Jon Pehrsson's children. In the following, however, the focus is only on the five of the triangulations that involved three of Jon Pehrsson's children: those on chromosome 1, 6, 12, 13 and 16. Because the strongest evidence of the hypothesis was found on chromosome 16, this is analysed first.

As an additional check of the Jon Pehrsson hypothesis, the possibility was considered that the important genealogical junction might lie one or more generations back in time. This was not relevant with regards to the genealogical junctions 1-3, because of their interconnectedness and their forward convergence in the Norberg family. The fourth genealogical junction, Jon Pehrsson (*1795) and his wife, however, was a "singleton" in this regard; here there was no interconnectedness with other genealogical junctions that contributed to the confirmation of the hypothesis. This, for example, left the possibility that it was not Jon Pehrsson who was the important genealogical junction but rather his parents. A search was therefore also made for DNA matches where his parents represented the MRCA, to see if this presented a viable alternative hypothesis.

3.5.2. Segment triangulation on chromosome 16

The strongest evidence in support of the hypothesis was found on chromosome 16, and more specifically on Brother 2's version of this chromosome. This was a chromosome where the two brothers shared very little DNA with each other. Brother 1, who shared only 15 cM of the segment at its very end, in fact had inherited almost all DNA on his corresponding chromosome 16 from the FF branch of the family (see above in section 3.4.4).

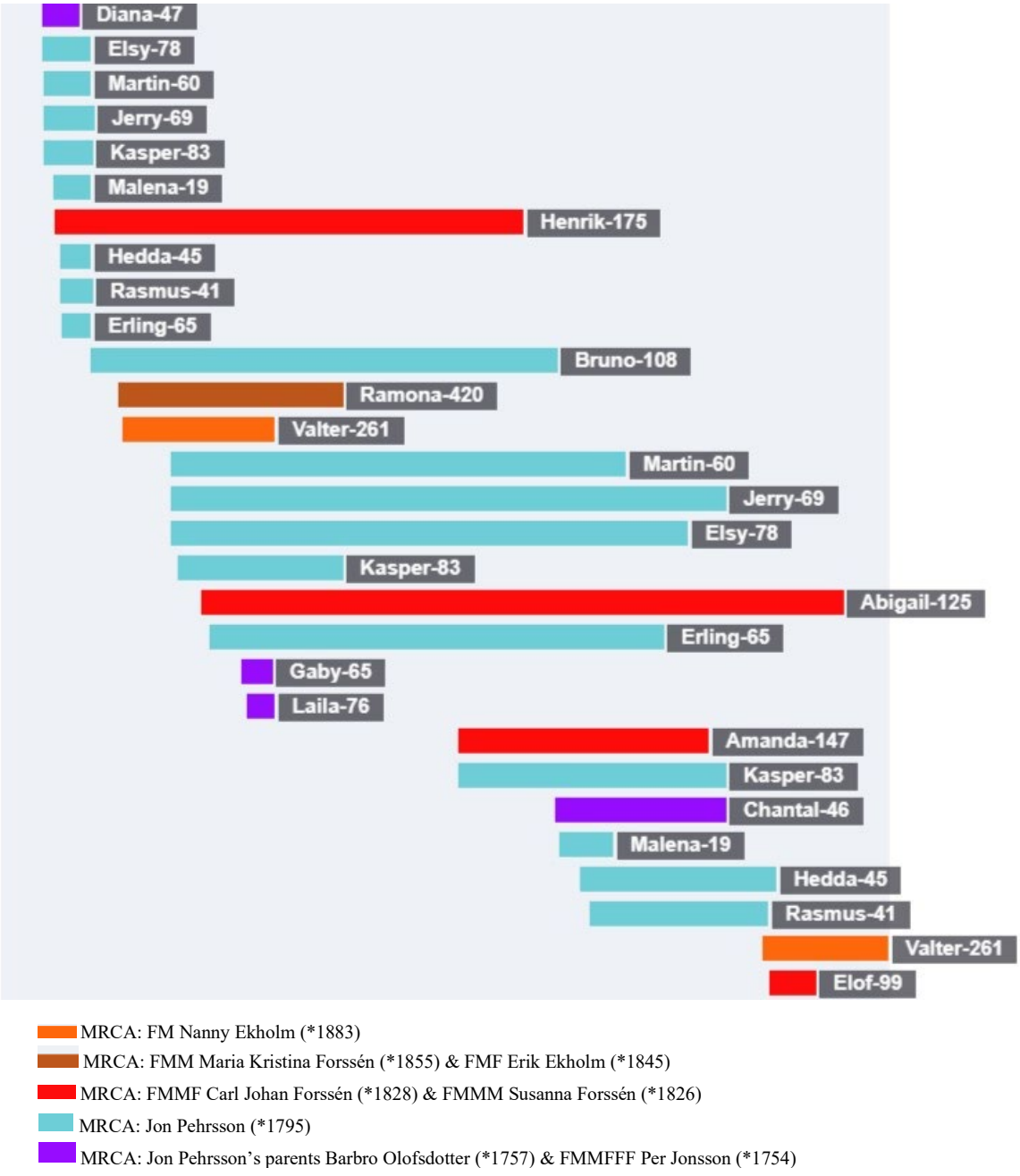
As seen in Table 6 and as illustrated in Figure 4, overlaps of DNA with six of Carl Johan Forssén's (*1828) descendants were found to *also* overlap with DNA of nine of Jon Pehrsson's (*1795) descendants on chromosome 16. Figure 4 shows the DNA segments shared with descendants of Jon Pehrsson (*1795) in blue colour (and his parents in purple), and segments shared with descendants of Carl Johan Forssén (*1828) in red (third cousins), brown (second cousin) or orange colour (first cousin level). As can be seen, the overlaps with these two groups of DNA matches cover a major part of chromosome 16, more specifically from location 4,171,701 to location location 78,883,118 (see also Table 6). The only two bits that are not covered by DNA segments that could be linked to descendants of *both* Jon Pehrsson and Carl Johan is the very first short part of the chromosome, from location 0 to location 4,171,701, and the very last part of the chromosome (from location 78,883,118 to the end of the chromosome). Although one of Carl Johan Forssén's descendants, *Valter-261*, shared this latter part of the chromosome with Brother 2 (and the very last part of it also with Brother 1), there was no evidence that any of Jon Pehrsson's children did.

The figure also shows some DNA matches with additional descendants of Jon Pehrsson's *parents* (*Diana-47, Gaby-65, Laila-76,* and *Chantal-46*) who shared small parts of these DNA segments on chromosome 16. They all turned out to be descendants of one of Jon Pehrsson's sisters. The relatively small number of these DNA matches, however, together with the relatively small DNA segments that were shared with them, gave no support for the possibility that his parents represented the most relevant genealogical junction.

The requirement of independent lineages among the DNA matches who shared segments on chromosome 16 was very well met. The nine descendants of Jon Pehrsson descended from three of his children (cf. table 6):

*Vilhelm Petter *1814* [*Hedda-45, Rasmus-41*]
*Erik *1820* [*Kasper-83, Elsy-78, Jerry-69, Erling-65, Martin-60, Malena-19*]
*Johan *1835* [*Bruno-108*]

Figure 4.
DNA segments on chromosome 16 (from location 4,171,701 to location 90,233,487) shared between Brother 2 and descendants of Jon Pehrsson (in blue) and his siblings (in purple) and descendants of Carl Johan Forssén (in shades of red, orange and brown depending on their most recent common ancestors [MRCA]), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage.



The six descendants of Carl Johan Forssén similarly could be divided into independent lineages via four of his children (cf. table 5):

*Maria Kristina *1855* [*Ramona-420, Valter-261*]
*Susanna Sofia *1857* [*Henrik-175, Amanda-147*]
*Karl *1860* [*Abigail-125*]
*Johan Petter *1862* [*Elof-99*]

The hypothesis was further strengthened by the fact that this segment triangulation involved intermediate MRCAs (Bartlett, 2016) covering three generation before reaching Jon Pehrsson (*1795) at generation 5:

Generation 2 [*Valter-261*; MRCA the author's FM]
Generation 3 [*Ramona-420*; MRCA the author's FMF/FMM]
Generation 4 [*Henrik-175, Amanda-147, Abigail-125, Elof-99*; MRCA the author's FMMF/FMMM]

This makes it possible to construct a hypothetical history of the major part of this chromosome from location 4,171,701 to location 78,883,118, as Brother 2 shared DNA *both* with descendants of Jon Pehrsson (*1795) and with descendants of Carl Johan Forssén (*1828) on this entire part of the chromosome. If this hypothesis were correct it would mean that the major part of this chromosome was derived from Jon Pehrsson (*1795) and had been transferred to his son Carl Johan Forssén (*1828), and from him to his daughter Maria Kristina Forssén (*1855), and in turn to her daughter Nanny Elina Ekholm (*1883), to her son (the author's F), and to his son Brother 2. Almost none of this DNA, however, was transferred to Brother 1, who instead received almost the whole of his corresponding chromosome 16 from the author's FF (see above in section 3.4.4).

In other words, if this were true it would mean that a large DNA segment covering a major part of this chromosome had been transferred relatively intact during five generations from Jon Pehrsson (*1795) to Brother 2 but had been completely lost to Brother 1 in the last stage of this process. Although Brother 1 shared 15 cM with Brother 2 on this chromosome, this small segment was located at the very end of the chromosome (from location 85,206,943 to 90,233,487) and although it was shared with *Valter-261* it showed no evidence of being shared with any descendant of Jon Pehrsson.

Although these results are entirely in line with the hypothesis, it is important to note that they do not in any way *prove* that Jon Pehrsson was the father of Carl Johan Forssén. For example, the results of the segment triangulation as such are equally theoretically compatible with the possibility that Jon Pehrsson was the father of Carl Johan's wife Susanna Forssén (*1826). All six descendants of Carl Johan Forssén that were presented in Figure 4 were of course, also descendants of Susanna Forssén. The DNA data presented in Table 4 can in no way differentiate between the hypothesis that Jon Pehrsson was the father of Carl Johan and the alternative hypothesis that he was the father of Susanna. *In combination with other data*, however, it appears that the Carl Johan hypothesis is much more likely to be true than the Susanna hypothesis. Most importantly, (1) Carl Johan's father was unknown, whereas Susanna's father is given in the population registers as Erik Eriksson Forssén (*1786); (2) Jon Pehrsson was a close neighbour to Carl Johan's mother in Pengsjö at the time when she got pregnant, whereas Susanna was from another village, Överboda, about ten kilometers to the east of Pengsjö. These demographic data make it much more *likely* that Jon Pehrsson was the father of Carl Johan than the father of Susanna.

Before leaving the analysis of the DNA segment at chromosome 16 it may be noted that of the three originally unplaced DNA matches, *Cesar-128, Bruno-108* and *Bianca-102,* who had Jon Pehrsson (*1795) as their ancestor (see Table 1 above, and section 3.2.4), only one (*Bruno-108*) shared DNA on chromosome 16 (a large segment of 51 cM). *Cesar-128* and *Bianca-102,* on the other hand, shared large segments with both brothers on

chromosome 1 and with Brother 1 on chromosome 12. The segment triangulations on these two chromosomes are described in the next two sections. Because there was some evidence that the DNA segment on chromosome 16 could be traced further back in time to Jon Pehrsson's parents, the family trees of other DNA matches who shared part of the segments on chromosomes 1 and 12 were also studied to see if they included close ancestors to Jon Pehrsson (*1795).

3.5.3. Segment triangulation on chromosome 1

Figure 5 depicts a large part of chromosome 1, from location 88,794,011 to the end of the chromosome. On this part of the chromosome, DNA segments were shared with two of Carl Johan Forssén's descendants (*Ramona-420* and *Robin-219;* marked in brown) and with several descendants of Jon Pehrsson (including the two strongest DNA matches, *Cesar-128* and *Bianca-102;* marked in blue), and also with some descendants of Jon Pehrsson's close maternal ancestors (marked in other shades of blue). In total, 16 descendants of Jon Pehrsson were found who shared segments on this part of the chromosome. Segment triangulations in support of the hypothesis were found for seven of them, including *Bianca-102,* who shared a large DNA-segment (62 cM) with brothers 1 and 2, parts of which were also shared with *Ramona-420* and *Robin-219* (see also Table 6). For *Cesar-128* and eight other Jon Pehrsson-descendants in the right part of the figure, however, no Carl Johan-descendants could be found that shared DNA with them on that part of the chromosome.

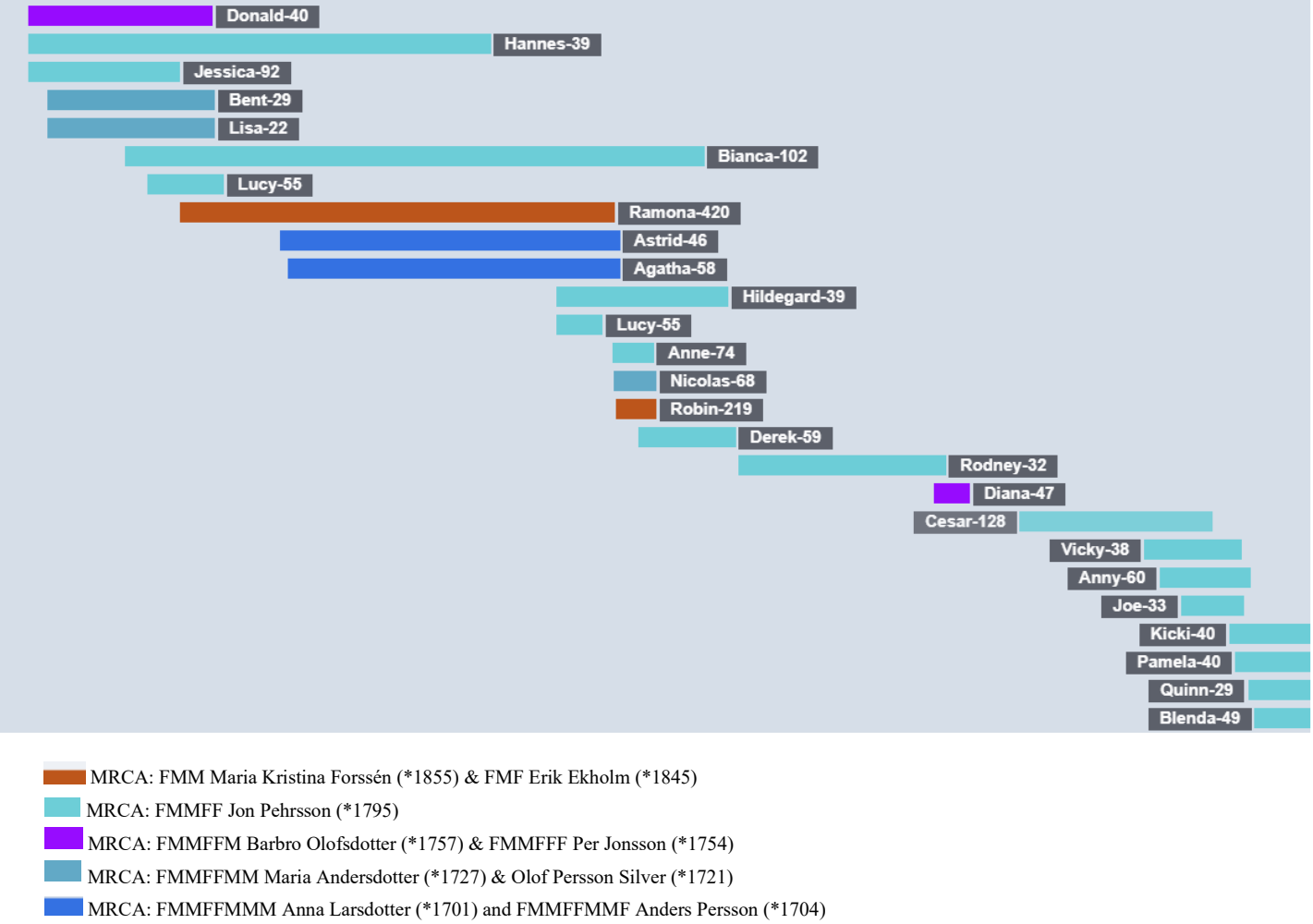
This means that, although the segment triangulation on chromosome 1 is clearly in line with the hypothesis, it does not provide equally strong evidence. Only two descendants of Carl Johan (*Ramona-420* and *Robin-219*) were found that shared the segment with *Bianca-102* and the six other Jon Pehrsson-descendants in the left part of the figure, and they were both related to the author at the second cousin level. No third cousin descendants of Carl Johan Forssén were found who shared DNA on this part of the chromosome. This leaves a gap of one generation in the analysis of this DNA segment, which theoretically makes room for other alternative hypotheses, as for example that Jon Pehrsson was an ancestor of the author's FMF Erik Ekholm (*1845). In other words, *the range of alternative hypotheses* are larger here than for the DNA segment on chromosome 16 (which was shared by four matches at the third cousin level).

In combination with other kinds of data (e.g., Jon Pehrsson being a neighbor to Carl Johan Forssén's mother in Pengsjö at the time she got pregnant), these alternative hypotheses still appear much less likely. Also, despite this increased uncertainty about how to interpret the segment triangulation in this case, the picture given in Figure 5 suggests the possibility that the entire chromosome from location 88,794,011 (where the segment shared with *Hannes-39* starts; see Table 6) to its end might derive from Jon Pehrsson (*1795). Sixteen of his descendants were found that "cover" that entire part of the chromosome, and Jon Pehrsson was the MRCA in all these cases.

The condition of independent lineages was well met. The seven descendants of Jon Pehrsson (*1795) that were part of the segment triangulation (in the left part of Figure 5) descended from three different children of his (cf. Table 6):

*Vilhelm Petter *1814* [*Derek-59, Lucy-55, Hannes-39, Hildegard-39*]
*Erik *1820* [*Anne-74*]
*Johan *1835* [*Bianca-102, Jessica-92*]

Figure 5.
DNA segments on chromosome 1 that the two brothers shared with descendants of Carl Johan Forssén (in red) and/or with descendants of Jon Pehrsson's children and his maternal ancestors (in various shades of blue depending on their most recent common ancestors [MRCA]), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage.



Some evidence also suggested that at least part of this DNA segment could be traced further back in time. As seen in Figure 5, Jon Pehrssons parents Barbro Olofsdotter (*1757) and Per Jonsson (*1754) were found to be the MRCAs of two DNA matches who shared part of the segment: *Donald-40* and *Diana-47*. Also, one generation further back in time, John Persson's maternal grandparents Maria Andersdotter (*1727) and Olof Persson Silver (*1721) were found to be the MRCAs of two additional DNA matches: *Astrid-46* and *Agatha-58*. And yet another generation back in time, two other matches were found to have Jon Pehrsson's mother's maternal grandparents Anna Larsdotter (*1701) and FMMFFMMF Anders Persson (*1704) in their pedigrees: *Bent-29* and *Lisa-22*. This is not only consistent with the hypothesis that the DNA segment on chromosome 1 derived from Jon Pehrsson (*1795), but it also suggests that it might be quite possible to trace DNA segments many generations back in time, at least in some cases.

Further corroboration of this possibility was found when a systematic search was engaged in for DNA matches who shared more than 15 cM of the segment from location 88,129,038 to location 146,672,906 on chromosome

1. The reason for engaging in this "side project" was that some rather strong DNA matches who bordered on the category of "closest unplaced DNA matches" (see section 3.1) shared rather large DNA segments with brothers 1 and 2 on this part of the chromosome. Most notably, *Filip-95* shared 38 cM of this segment. A search of his pedigree showed that the MRCA with *Filip-95* seemed to be Anna Larsdotter's (*1701) parents Märeta Mattsdotter *(1678 in Armsjö, Nordmaling) and Lars Olofsson (*1675 in Bergsjö, Nordmaling). This couple was Jon Pehrsson's (*1795) MMMM and MMMF, and in other words nine generations back in time from the author.

Surprisingly, a systematic search for other DNA matches who shared more than 15 cM of the segment shared with *Filip-95* led to the identification of 43 additional matches. Of these 32 had Märeta Mattsdotter (*1678) and Lars Olofsson (*1675) in their pedigrees (in most cases as the MRCA), and ten further matches had Märeta Mattsdotter's (*1675) father Matts Hindersson (*around 1618)[3]. Only one of the 43 matches fell outside the picture, as no CA with her could be found (possibly because her mother's paternal grandfather was unknown). Figure 6 describes these DNA segments. This figure represents an expansion of Figure 5 with a focus on the specific region of chromosome 1 from location 88,129,038 to location 146,672,906, and including DNA matches with MRCAs two more generations back in time.

Typically, the size of DNA segments shared with *closer* cousins tend to be larger than the segments shared with more *distant* cousins. But here the size of the segments shared with descendants of Märeta Mattsdotter (*1678) and her husband tended to be *larger* than those shared with her daughter Anna Larsdotter (*1701), her granddaughter Maria Andersdotter (*1727), and her great granddaughter Barbro Olofsdotter (*1757). So many as 12 of the 32 DNA matches where Märeta and her husband appeared as MRCA were larger than 30 cM. Furthermore, many more DNA matches were found who shared this DNA segment with Märeta Mattsdotter than with her descendants in Jon Pehrsson's (*1795) maternal ancestral line: 32 matches as compared with only a few in each generation after her. This raises several questions for further research.

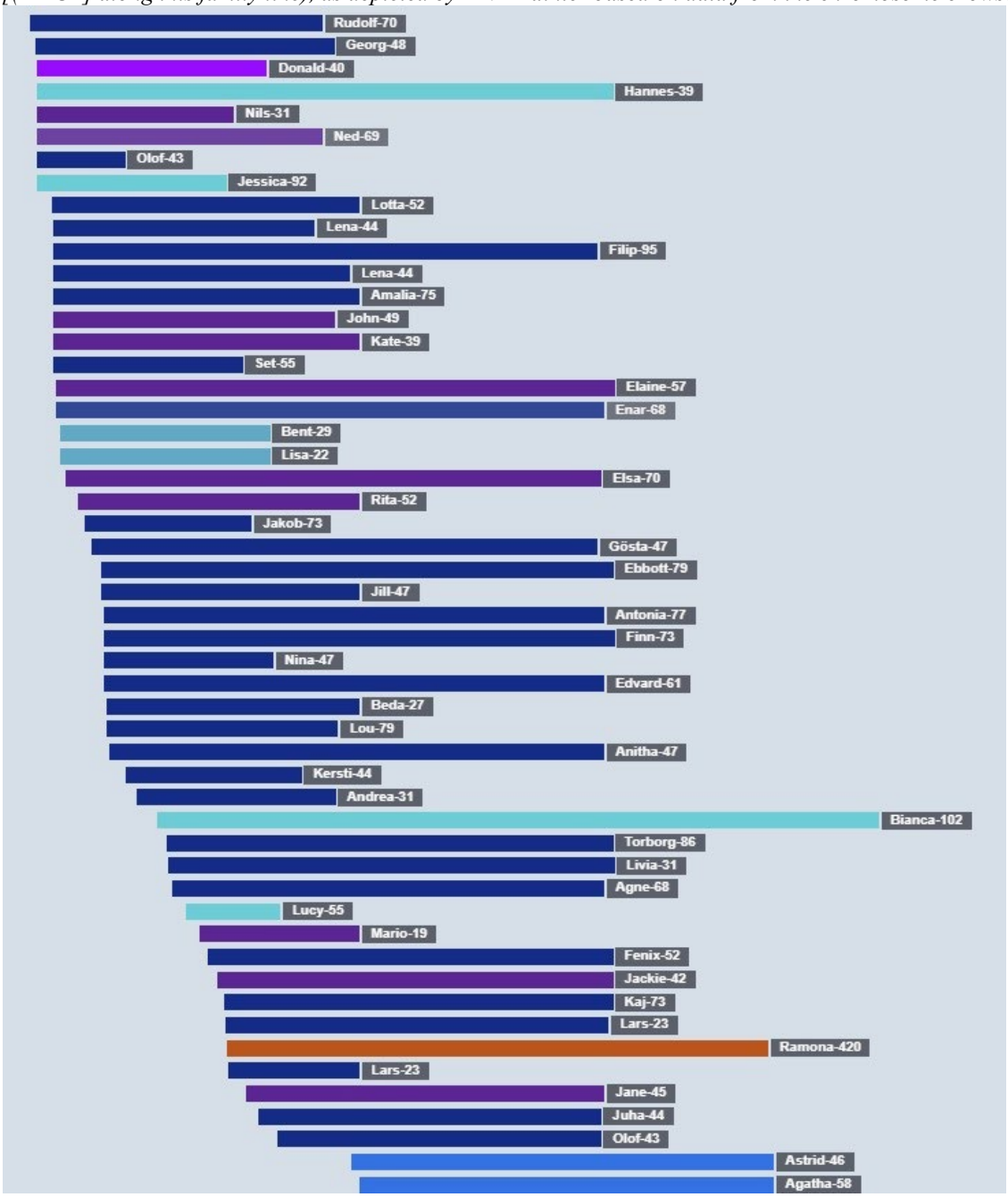
One possible contributing factor would be pedigree collapse among the DNA matches. Examples of this were found among several of these DNA matches. To take one of the more extreme examples: *Ebbott-79* shared 33 cM of this DNA segment, and Märeta Mattsdotter (*1678) and Lars Olofsson (*1675) were the most recent common ancestors (MRCA) that could be found in his pedigree. Moreover, this couple was found in nine places (eight generations back in four cases, and nine generations back in five cases) in his pedigree. In terms of genealogical relationships with *Ebbot-79* this means that he was a 4 x 7th cousin-1R and 5 x 8th cousin via Märeta Mattsdotter and her husband. In addition, Märeta's siblings and half siblings were found in 14 more places in his pedigree, which means that *Ebbott-79* was also the author's cousin via Matts Hindersson (*1618) in fourteen additional cases. This might have contributed to the amount of DNA shared with *Ebbot-79* both on chromosome 1 (33 cM) and in total (79 cM).

This could hardly be the full explanation of the present findings, however, because several other of these DNA matches apparently had only one lineage to Märeta Mattsdotter and her husband. For example, *Filip-95* who was the DNA match in this group who shared the largest amount of DNA with the author and his brother, apparently had Märeta Mattsdotter and her husband only at one place in his pedigree.

[3] Only her father, not her mother, is mentioned here because the lineages from Matts Hindersson (*1618) involved children from his two marriages.

Figure 6.

DNA segments on chromosome 1 that Brothers 1 and 2 shared with descendants of Carl Johan Forssén (in brown) and descendants of Jon Pehrsson's children and his maternal ancestors (in various shades of blue depending on their most recent common ancestors [(MRCA] along this family line), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage



MRCA: FMM Maria Kristina Forssén (*1855) & FMF Erik Ekholm (*1845)
MRCA: FMMFF Jon Pehrsson (*1795)
MRCA: FMMFFM Barbro Olofsdotter (*1757) & FMMFFF Per Jonsson (*1754)
MRCA: FMMFFMM Maria Andersdotter (*1727) & Olof Persson Silver (*1721)
MRCA: FMMFFMMM Anna Larsdotter (*1701) & FMMFFMMF Anders Persson (*1704)
MRCA: FMMFFMMMM Märeta Mattsdotter (*1678) & FMMFFMMMF Lars Olofsson (*1675)
CA: FMMFFMMMMF Matts Hindersson (*around 1618)

Another possibility in terms of pedigree collapse would be that Märeta Mattsdotter (*1678) and Lars Olofsson (*1675) were ancestors to the author also via some additional linages/lineages. Although this couple could not be found anywhere else in the author's pedigree, Märeta's half-sister Elisabet Mattsdotter (*1661) was found in his FF branch as the FF's FMFFMFM, which means that their father Matts Hindersson (*around 1618) was an ancestor both to the author's FF and his FM at ten generations distance.

Perhaps even more relevant here was the possibility that Märeta Mattsdotter and her husband might have an additional place somewhere in the author's pedigree, via unknown ancestors further back in time. Most notably, the father of Carl Johan Forssén's (*1828) MMF Hans Jonsson Tiger (*1742) was unknown, as were also the parents of his wife Magdalena Eriksdotter (*around 1742). The possibility of a connection here was suggested by the fact that three of the DNA matches in Figure 6 (*Elaine-57, John-49*, and *Kate-39*) had the Hans/Magdalena couple as MRCA in their pedigrees. They are depicted in the figure as having Matts Hindersson (*around 1618) as their CA (*not* MRCA), but in view of the relatively large segments they shared (37 cM, 31 cM, and 30 cM, respectively) it seemed more likely that the explanation might go via their MRCA, who were three generations closer in time. If so, however, this would require some genealogical connection between the Hans/Magdalena couple and the Märeta/Lars couple.

3.5.4. Segment triangulation on chromosome 12

Figure 7.
DNA segments on chromosome 12 that Brother 1 shared with descendants of Jon Pehrsson's children and his maternal ancestors (in various shades of blue depending on their most recent common ancestors [MRCA]) and descendants of Carl Johan Forssén (in brown), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage.

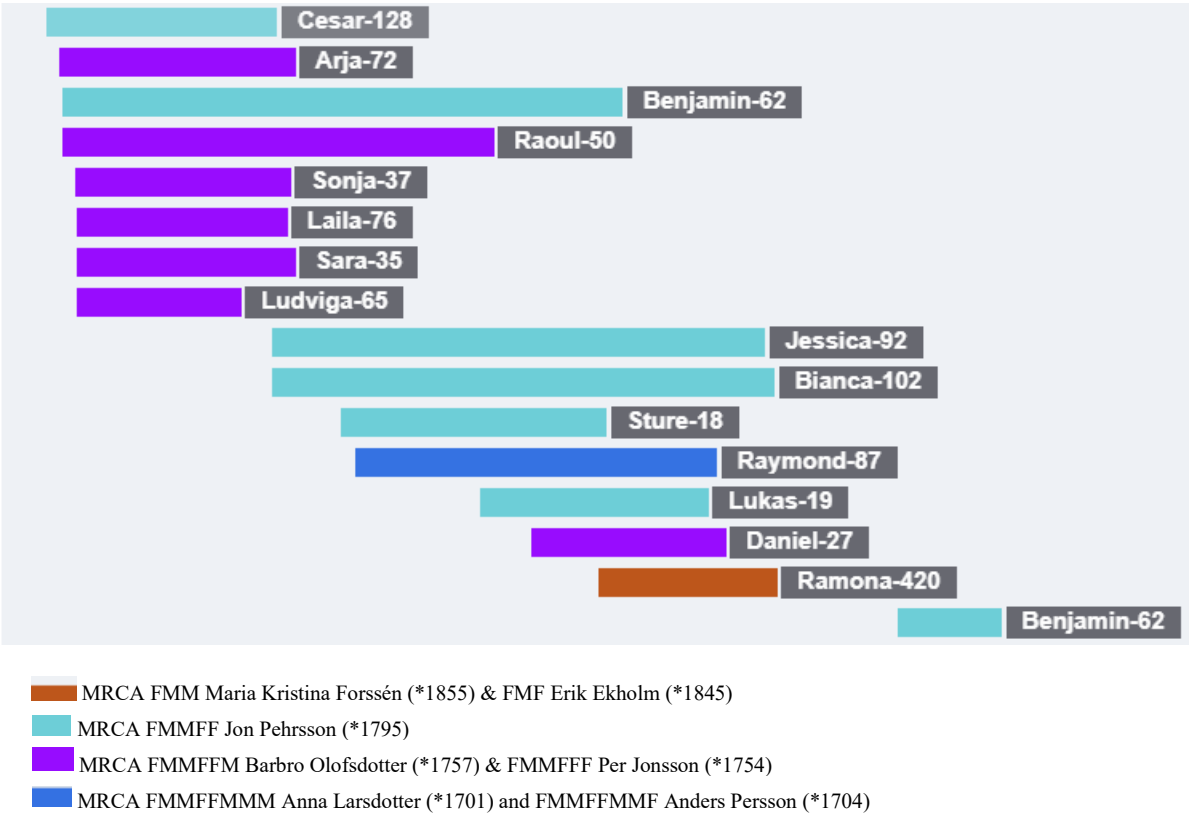


Figure 7 describes the segment triangulation on chromosome 12. As shown in the figure, this segment triangulation involved only one descendant of Carl Johan Forssén (*1828), *Ramona-420,* who shared a relatively small part of the segment (17 cM). This was in stark contrast to the relatively large segments shared with descendants of Jon Pehrsson (*1795): 55 cM with *Benjamin-62*, 40 cM with *Bianca-102*, 38 cM with *Jessica-92*, and 34 cM with *Cesar-128*.

Yet the condition of independent lineages was well met also here; the six descendants of Jon Pehrsson (*1795) who were part of the segment triangulation descended from three of his children (cf. Table 6):

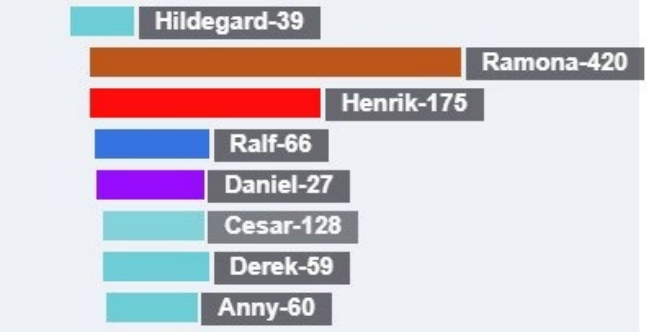
*Erik *1820* [*Cesar-128, Benjamin-62, Sture-18*]
*Anna Margareta *1821* [*Lukas-19*]
*Johan *1835* [*Bianca-102, Jessica-92*]

As seen in Figure 7, the triangulations also involved several DNA matches who had Jon Pehrsson's parents and maternal grandparents as their MRCAs, which was consistent with the hypothesis. On the other hand, the condition of intermediate MRCAs was not met (no 3rd cousins were found who shared parts of the segment), which means that this segment triangulation considered in isolation represented rather weak evidence of the hypothesis.

3.5.5. Segment triangulation on chromosome 6

Figure 8 shows the segment triangulation on chromosome 8. Here the condition of intermediate MRCAs was met, as the segment was shared by both second and third cousins.

Figure 8.
DNA segments on chromosome 6 that Brother 1 shared with descendants of Jon Pehrsson and his maternal ancestors (in various shades of blue depending on their most recent common ancestors (MRCA) and with descendants of Carl Johan Forssén (in red and brown), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage.



MRCA: FMM Maria Kristina Forssén (*1855) & FMF Erik Ekholm (*1845)
MRCA: FMMF Carl Johan Forssén (*1828) & FMMM Susanna Forssén (*1826)
MRCA: FMMFF Jon Pehrsson (*1795)
MRCA: FMMFFM Barbro Olofsdotter (*1757) & FMMFFF Per Jonsson (*1754)
MRCA: FMMFFMMM Anna Larsdotter (*1701) and FMMFFMMF Anders Persson (*1704)

As to the segment triangulation on chromosome 6 (see Figure 8) it also met the condition of independent lineages quite well. The four descendants of Jon Pehrsson (*1795) who were part of the segment triangulation descended from three different children of his (cf. Table 6):

*Erik *1820* [*Cesar-128*]
*Vilhelm Petter *1814* [*Hildegard-39, Derek-59*]
*Johan *1835* [*Anny-60*]

Further, some DNA matches with Jon Pehrsson's (*1795) maternal ancestors were also found to share DNA on this segment, which was clearly in line with the hypothesis.

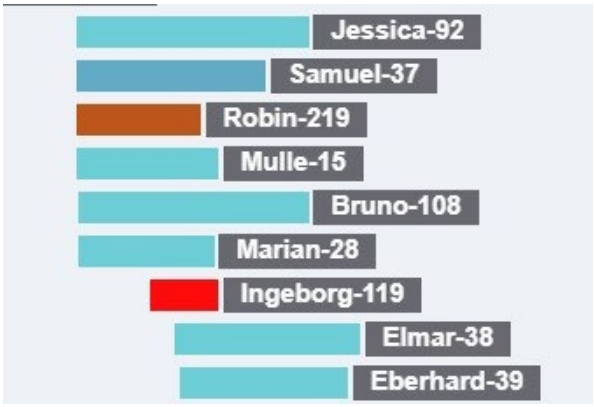
3.5.5. Segment triangulation on chromosome 13

Finally, as to the segment triangulation on chromosome 13, the situation was rather similar to that on chromosome 6. The condition of independent lineages was well met as the six descendants of Jon Pehrsson (*1795) who were part of the segment triangulation descended from three different children (cf. Table 6):

*Erik *1820* [*Elmar-38, Eberhard-39*]
*Cajsa Lisa *1828* [*Mulle-15, Marian-28*]
*Johan *1835* [*Jessica-92, Bruno-108*]

The condition of intermediate MRCAs was also met. As seen in Figure 9, the triangulation involved descendants of Carl Johan Forssén (*1828) at both the second cousin and third cousin level.

Figure 9.
DNA segments on chromosome 13 that Brother 1 shared with descendants of Jon Pehrsson and his maternal ancestors (in various shades of blue depending on their most recent common ancestors [MRCA] and with descendants of Carl Johan Forssén (in red and brown), as depicted by DNA Painter based on data from the chromosome browser in MyHeritage.



MRCA: FM Nanny Ekholm (*1883)
MRCA: FMM Maria Kristina Forssén (*1855) & FMF Erik Ekholm (*1845)
MRCA: FMMF Carl Johan Forssén (*1828) & FMMM Susanna Forssén (*1826)
MRCA: FMMFF Jon Pehrsson (*1795)
MRCA: FMMFFMM Maria Andersdotter (*1727) & Olof Persson Silver (*1721)

To summarize, the segment triangulations contributed considerable evidence in support of the FMMFF Jon Pehrsson hypothesis. In combination with demographic data showing that Jon Pehrsson was a close neighbour to Greta Stina Olofsdotter (*1800) at the time when she got pregnant with her son Carl Johan (*1828), the large number of segment triangulations (and especially the triangulation on chromosome 16) presents a strong case for Jon Pehrsson being the father of Carl Johan.

3.6. Evaluation of the genealogical junctions method

In the present study, the analysis of genealogical junctions led to the identification of genealogical relationships to previously unplaced DNA matches, and to the filling of several gaps in the author's pedigree. But it did not lead to the identification of the genealogical relationships to *all* unplaced DNA matches, and it did not *directly* identify all ancestors that were searched for. The present section focuses on these apparent shortcomings, first with regards to the unplaced DNA matches, and then with regards to the newly identified ancestors.

3.6.1. The previously unplaced DNA matches

Table 1 above contained a list of 19 unplaced DNA matches. Table 7 describes the genealogical relationships that were discovered to these persons as results of the analysis.

Table 7.
The previously unplaced DNA matches of the two brothers, and genealogical relationships that were established to them.

Code namn	*Genealogical relationship*	*Most recent common ancestors (MRCA)*
Sofia-150	3rd cousin	Johan Petter Johansson (*1819) & Maria Karolina Mattsdotter (*1829)
Thore-147		
Eivor-144	3rd cousin	Erik Olofsson (*1782) & Anna Katarina Jakobsdotter (*1788)
Maja-132	3rd cousin 1R	Matts Olofsson (*1806) & Maja Greta Andersdotter (*1803)
Cesar-128	half 4th cousin	Jon Pehrsson (*1795)
Willy-124		
Igor-124	3rd cousin	Johan Petter Johansson (*1819) & Maria Karolina Mattsdotter (*1829)
George-122	4th cousin	Erik Olofsson (*1782) & Anna Katarina Jakobsdotter (*1788)
Austin-120	4th cousin	Erik Olofsson (*1782) & Anna Katarina Jakobsdotter (*1788)
Tora-116	4th cousin	Matts Olofsson (*1806) & Maja Greta Andersdotter (*1803)
Clara-114	3rd cousin 1R	Johan Petter Johansson (*1819) & Maria Karolina Mattsdotter (*1829)
Axel-112		
Elisabeth-115	4th cousin	Matts Olofsson (*1806) & Maja Greta Andersdotter (*1803)
Marcus-111	4th cousin	Matts Olofsson (*1806) & Maja Greta Andersdotter (*1803)
Bruno-108	half 3rd cousin 1R	Jon Pehrsson (*1795)
Ellie-106		
Bianca-102	half 5th cousin 1R	Jon Pehrsson (*1795)
Ludvig-101	4th cousin 1R	Matts Olofsson (*1806) & Maja Greta Andersdotter (*1803)
Elvira-101	4th cousin	Erik Olofsson (*1782) & Anna Katarina Jakobsdotter (*1788)

As can be seen in Table 7, fifteen of these 19 previously unplaced DNA matches were found to be either 3rd, 4th, or 5th cousins, one step removed at most. Four of them, however, remained unplaced: *Thore-147, Willy-124, Axel-112,* and *Ellie-106*. As to *Thore-147,* this was most probably due to his FFF being unknown. *Thore-147*'s FF and FFM, however, were from Nordmaling parish, and he shared considerable amounts of DNA with several other of the author's closest matches on the FF side: 102 cM with *Niklas-290*, 138 cM with *Maja-132*, 109 cM with *Ludvig-101*, and 154 cM with *Saga-91*. A reasonable hypothesis was that the relationship with *Thore-147* went

via his unknown FFF, which in that case probably would make him a 2nd cousin-1R, a 3rd cousin, or maybe a 3rd cousin-1R.

A closer investigation of the family trees of *Willy-124* and *Axel-112* also led to the identification of genealogical relationships that might explain the amount of DNA shared with them. As to *Willy-142*, he turned out to be related to the author in multiple ways. First, he was the author's 5th cousin one step removed via *the parents of* Anna Katarina Jakobsdotter (*1788) from GJ-2. Second, *Willy-142* was the author's fourfold 7th cousin via one couple from the author's FM branch (i.e., a case of pedigree collapse, as this couple appeared at four slots in *Willy-142's* family tree). A close inspection of the DNA segments that were shared with *Willy-142* revealed that the author shared DNA with him both via FF and FM. These multiple genealogical relationships seemed to be a reasonable explanation of the total amount of DNA that was shared.

As to *Axel-112*, he turned out to be the author's 5th cousin in two different ways: (1) *via the parents* of Erik Olofsson (*1782) from GJ-2, and (2) *via the maternal grandparents* of Johan Petter Johansson (*1819) from GJ-3. This double 5th cousin relationship seemed to be a reasonable explanation of the total amount of DNA that was shared. Because these relationships were at the level of 5th cousins they could, by definition, not be detected by the analysis of genealogical junctions as it was defined in the present study.

Finally, as to *Ellie-106*, it was more difficult to find a likely explanation of the amount of DNA shared with her. No genealogical relationships even at a 5th cousin level could be found. The search for an MRCA with *Ellie-106* led to Jon Pehrsson's maternal grandparents, Maria Andersdotter (*1727) and Olof Persson Silver (*1721). The largest DNA segment shared with her (33 cM on chromosome 20) was shared by one descendant of FMMF Carl Johan Forssén (*Abigail-125*) and another descendant of Jon Pehrsson (*Jörgen-68*), which suggested that at least part of the DNA shared with *Ellie-106* came from Jon Pehrsson's mother's branch of the family tree. More specifically this meant that *Ellie-106* was the author's sixth cousin one step removed, but this seemed too far back in time to be a sufficient explanation of all the amount of DNA shared. A close inspection of her family tree also revealed some examples of pedigree collapse; for example, she was a threefold 7th cousin via the author's FMMFMFFF/ FMMFMFFM, who were found at three slots in her family tree. This suggested the possibility that the DNA shared with her resulted from multiple genealogical relationships via the FM branch.

3.6.2. The newly identified ancestors

One main result was that the many empty slots in the author's pedigree, as shown in Figure 1, were now filled. The resulting new pedigree is seen in Figure 9.

Figure 9.
The author's father's pedigree, as known at the end of the investigation

TWO GENERATIONS BACK FROM THE AUTHOR
FF: One of the Norberg brothers
FM: Nanny Elina Ekholm *1883-11-10 Nordåker, Vännäs †1948-09-03 Robertsfors

THREE GENERATIONS BACK
FFF: Erik Olof Norberg *1850-10-31 Lögdeå, Nordmaling †1935-02-14 Nordmaling
FFM: Klara Maria Johansdotter *1858-04-25 Ava, Nordmaling †1935-01-20 Lögdeå, Nordmaling
FMF: Erik Ekholm *1845-12-12 Östanå, Vännäs †1920-11-13 Hörneå, Hörnefors
FMM: Maria Kristina Forssén *1855-03-16 Överboda, Umeå †1915-09-16 Hörneå, Hörnefors

FOUR GENERATIONS BACK
FFFF: Olof Norberg Jonsson *1819-02-03 Pengsjö, Vännäs †1892-03-22 Lögdeå, Nordmaling

FFFM: Johanna Eriksdotter *1824-05-25 Mullsjö, Nordmaling †1854-03-20 Lögdeå, Nordmaling
FFMF: Johan Petter Johansson *1819-11-26 Ava, Nordmaling †1908-09-03 Ava, Nordmaling
FFMM: Maria Karolina Mattsdotter *1829-05-29 Ava, Nordmaling †1914-02-12 Ava, Nordmaling
FMFF: Johan Georg Ekholm *1821-06-13 Gumboda, Nysätra †1905-02-05 Nordåker, Vännäs
FMFM: Anna Anna Elisabet Lundberg *1820-09-19 Norrmjöle, Umeå †1886-11-18 Nordåker, Vännäs
FMMF: Carl Johan Forssén *1828-02-15 Björnlandsbäck, Vännäs †1865-06-08 Granlund, Vännäs
FMMM: Susanna Forssén *1826-06-24 Överboda, Umeå †1897-02-01 Granlund, Vännäs

FIVE GENERATIONS BACK
FFFFF: Jon Jonsson *1781-11-09 Pengsjö, Vännäs †1860-03-15 Pengsjö, Vännäs
FFFFM: Magdalena Eriksdotter Vänman *1786-07-20 Vännäs †1870-03-18 Pengsjö, Vännäs
FFFMF: Erik Olofsson *1782-09-05 Mullsjö, Nordmaling †1846-05-20 Mullsjö, Nordmaling
FFFMM: Anna Katarina Jakobsdotter *1788-03-24 Örsbäck, Nordmaling †1864-07-04 Mullsjö, Nordmaling
FFMFF: Johan Persson *1789-09-26 Gideåbacka, Grundsunda †1868-07-30 Ava, Nordmaling
FFMFM: Kristina Nilsdotter *1792-09-07 Rönnholm, Nordmaling †1855 Ava, Nordmaling
FFMMF: Matts Olofsson *1806-08-30 Bodum, Grundsunda †1863-06-18 Bodum
FFMMM: Maja Greta Andersdotter *1803-08-25 Långed, Nordmaling †1887-09-12 Ava, Nordmaling
FMFFF: Göran Johansson Ekholm *1789-12-31 Forsa †1859-04-14 Östanå, Vännäs
FMFFM: Margareta Larsdotter *1779-01-14 Gumboda, Nysätra †1841-07-29 Östanå, Vännäs
FMFMF: Fredrik Lundberg *1777-08-12 Korsholm, Vasa, Finland † 1837-10-19 Norrmjöle, Umeå
FMFMF: Lisa Greta Persdotter *1779-09-14 Gubböle 2, Umeå † 1851-02-14 Norrmjöle, Umeå
FMMFF: Jon Pehrsson *1795-09-24 Berg, Vännäs †1889-03-03 Brännland, Bjurholm
FMMFM: Greta Stina Olofsdotter *1800-01-29 Hjuken, Vindeln †1890-09-11 Nylandsnäs, Vännäs
FMMMF: Erik Eriksson Forssén *1786-02-07 Nyåker, Nordmaling †1867-12-12 Överboda, Umeå
FMMMM: Charlotta Andersdotter *1790-05-22 Överboda, Umeå

Focusing on the sixteen ancestors five generations back in time, as seen in Figure 9, only seven of them were known before this study (see Figure 1), whereas nine were missing. Of these nine previously unknown ancestors seven appeared in the analysis of genealogical junctions: FFFMF/FFFMM Erik Olofsson (*1782) and Anna-Karin Jakobsdotter (*1788) in GJ-2; FFMFF/FFMFM Johan Persson (*1789) and Kristina Nilsdotter (*1792) as being the parents of the Johan Petter Johansson (*1819) in GJ-3; FFMMF/FFMMM Matts Olofsson (*1806) and Maja Greta Andersdotter (*1803) in GJ-1; and FMMFF Jon Pehrsson (*1795) in GJ-4. The fact remains, however, that one of the couples at the distance of 5 generations was *not* found in the analysis of genealogical junctions: the couple FFFFF/FFFFM Jon Jonsson (*1781) and Magdalena Eriksdotter Vänman (*1786).

This gives rise to some questions of relevance for understanding the reliability and sensitivity of the analysis of genealogical junctions. A first question is why the FFFFF/FFFFM couple was not found in this analysis. A second question is if they could have been found with a lower cut-off for including unplaced DNA matches (i.e., lower than 100 cM).

One possible explanation why a certain couple is not found in this kind of analysis is that they have fewer descendants that could test their DNA and is therefore less likely to appear in the family trees of DNA matches. This, however, did not seem to the case here. The couple FFFFF/FFFFM Jon Jonsson (*1781) and Magdalena Eriksdotter Vänman (*1786) had eight children, of which seven formed their own families, and altogether they had 62 grandchildren. This was not lower than those of the couples in genealogical junctions 1-4 (see section 3.2).

An exploration of the family trees of the unplaced DNA matches showed that only one of them, *Eivor-144*, did have the FFFFF/FFFFM couple Jon Jonsson (*1781) and Magdalena Eriksdotter Vänman (*1786) in her family tree. This did not pass the cut-off for genealogical junctions, which required that at least two unplaced DNA matches should have the couple in their family trees. In fact, the cut-off would have to be lowered all the way to 86 cM to make this couple a genealogical junction. With such a lowered cut-off, *Lola-86* had also been included among the unplaced DNA matches, as she had the FFFFF/FFFFM couple at two different places in her family tree.

On the other hand, a lowering of the cut-off to 86 cM would probably also have led to the identification of other genealogical junctions that might not be relevant to the author's family tree. In other words, it would probably have led to false positives (i.e., the identification of couples who were common ancestors to two or more unplaced DNA matches without having any place in the author's family tree).

Finally, the method of analyzing genealogical junctions was also tested by exploring if it could have identified ancestors on the FM side, if that had been the task. This was clearly so in the case of the FM's mother's family, as eight of the closest DNA matches (with a cut-off of 100 cM) had FMMF Carl Johan Forssén (*1828) and FMMM Susanna Forssén (*1826) in their family trees. This was not equally clear, however, with the FM's father's family. Two of the closest DNA matches (>100 cM), *Linnea-112* and *Olivia-105*, did have the FMFMF/FMFMM couple Fredrik Lundberg (*1777) and Lisa Greta Persdotter (*1779) in their family trees. But only one of them, *Olivia-105*, also had the FMFFF/FMFFM couple Göran Johansson Ekholm (*1789) and Margareta Larsdotter (*1779) in her tree.

This means that the latter couple would not have been identified by an analysis of genealogical junctions. Moreover, it would not have helped to lower the cut-off in this case, as the author was not able to find any additional DNA match who had this couple in their family tree even with as low a cut-off as 50 cM. One possible explanation was that the FMFFF/FMFFM couple had fewer living descendants that could test their DNA. In fact, they had six children, of which only three grew up to form their own families, and altogether 19 grandchildren. This was well below all the other ancestors at the same generational distance (see section 3.2).

4. DISCUSSION

The analysis of genealogical junctions proved to be a successful method in the present study. Not only did it lead to the identification of the family of origin of the author's paternal grandfather (the Norberg family) but also to the identification of a person who most probably was the author's FMMFF: Jon Pehrsson (*1795). Admittedly, the analysis failed to identify specifically who among the Norberg brothers was the FF, but this was not due to any weakness of the method but to the fact that three of the FF candidates did not have any known children and therefore no living descendants who could test their DNA.

4.1 Evidence and proof

Genealogists sometimes ask for "proof" in these matters. *Proofs*, however, belong to mathematics and formal logic and not to empirical sciences such as genetic genealogy. What matters in the empirical sciences is *evidence*. Evidence may vary in strength, but never attains the status of proofs in the strict sense of the word, because it is always possible to imagine alternative hypotheses that are logically possible even when they are extremely unlikely. In the present study, the evidence for the FF Norberg hypothesis was quite strong. The Norberg hypothesis was supported by (1) the interconnectedness of three different genealogical junctions (GJ1, GJ2, and GJ3) which covered most of the author's closest unplaced DNA matches and converged in one specific family, the Norberg family, (2) atDNA testing of four of five available Norberg descendants, (3) Y-DNA testing of one of the Norberg descendants, and (4) segment triangulation.

Still, it is possible to imagine alternative hypotheses that are compatible with the findings, although they may seem extremely unlikely. For example, it is logically possible that the mother in the Norberg family, Klara Maria Johansdotter (*1858), had an unknown sister (it must be an *unknown* sister, because according to the records she

had only two brothers, no sister) who for some reason was sent away to be raised in another family. Suppose that the father in the Norberg family, Erik Olof Norberg (*1850), met this unknown sister and they had a child together, and that this child grew up to be a man who was the real FF. This would explain all the available data equally well as the Norberg hypothesis, although it may seem to be an extremely unlikely hypothesis. It is probably possible to construct several other far-fetched logical possibilities. Furthermore, it cannot be completely ruled out that new data might be revealed that would change the picture and make an alternative hypothesis plausible.

In this perspective, it may be worth remembering that the success in this part of the study came after repeated failures to identify the FF in several other ways, based on suggestions from relatives to the foster family and attempts from representatives of the municipality, as well as the author's own search for possible candidates in the FM's vicinity. It needs to be emphasized that the success of the Norberg hypothesis was based *entirely* on genetic-genealogical data, as there was no other evidence even of a meeting between the author's FM and any of the Norberg brothers. No social connections were found between the FM and the Norberg family, and this family could probably never have been identified via any other sources. Yet the Norberg hypothesis was supported both by atDNA testing and a test of Y-DNA, which together represents very strong evidence in support of the hypothesis.

As to the identification of the FMMFF, the evidence is different. Here it is quite easy to find alternative hypotheses that are compatible with the genetic-genealogical data. For example, as already stated in section 3.5.2 on p. 27, the results of the segment triangulations are quite compatible with the possibility that Jon Pehrsson (*1795) was the father of the author's FMMM Susanna Forssén (*1826) rather than being the father of her husband FMMF Carl Johan (*1828). The DNA data from the segment triangulations cannot differentiate between these two hypotheses. *In combination with demographic data*, however, the Carl Johan hypothesis is much more likely to be true than the Susanna hypothesis. Most importantly, (1) Carl Johan's father was unknown, whereas this was not the case with Susanna's father; and (2) Jon Pehrsson was a close neighbour to Carl Johan's mother in Pengsjö at the time when she got pregnant, whereas Susanna was born in another village. These data make it much more *likely* that Jon Pehrsson was the father of Carl Johan than the father of Susanna.

The present results also illustrate the crucial importance of having one's siblings DNA tested. The most conclusive evidence of Jon Pehrsson (*1795) being the FMMFF was found on Brother 2's version of chromosome 16. On this chromosome, large convincing segment triangulations were found between nine descendants of Jon Pehrsson's children and six descendants of the author's FMMF Carl Johan Forssén (*1828). Moreover, these triangulations satisfied the conditions of independent lineages (i.e., going via three of Jon Pehrsson's other children) and the involvement of intermediate MRCAs at the level of 1st, 2nd, and 3rd cousins (cf. Bartlett, 2016; Thomas, 2021). Several triangulations on other chromosomes served to corroborate and strengthen the hypothesis even further.

4.2. The importance of the number of descendants

A more general consideration concerning the analysis of genealogical junctions is that it is clearly sensitive to the number of living descendants of a given ancestor. An illustration of this was found in the last part of the results section, where the method was tested on all the author's ancestors five generations back in time in the paternal branch of his family tree. Of these altogether sixteen ancestors, four (two couples) were not identifiable by the present kind of analysis of genealogical junctions. The method failed with regards to the FFFFF/FFFFM couple Jon Jonsson (*1781) and Magdalena Eriksdotter Vänman (*1786) and with regards to the FMFFF/FMFFM couple

Göran Johansson Ekholm (*1789) and Margareta Larsdotter (*1779). Each of these couples were found in only one family tree of the strongest DNA matches (i.e., those who passed the cut-off of >100 cM shared DNA).

As to the latter couple one part of the explanation most probably was that they had considerably fewer children and grandchildren and therefore fewer living descendants that could test their DNA. The following general principle seems to hold: *The fewer descendants an ancestor has, the fewer DNA matches will probably be found who has this ancestor in their family tree.* It is even quite possible that one shares considerable amounts of atDNA with a certain ancestor five generations back in time without this ever being *possible* to detect via atDNA testing. To see this one may consider the following thought experiment:

Suppose you have an ancestor five generations back in time, Carl Smith, who was born in 1795. Suppose further that this ancestor had several children, but only one of them, Jim Smith (*1828), survived to raise a family. Suppose further that of Jim's children in turn only two of them had children of their own, but only one of these, Carla Smith (*1855), had children of her own, and that only one of them in turn had children: a woman by the name of Anna Smith (*1883). Suppose that Anna Smith is your grandmother, and that she also had only one child: your father. In that case Carl Smith (*1795) would be your FMMFF, but he would have no other living descendants that could possibly test their DNA; you would be his only living descendant. In other words, he could not possibly be identified by any genetic-genealogical method.

Although this thought experiment may describe an extreme case that is probably rarely seen in the real world, such limiting cases may be important to establish general principles. And although such an extreme case may be rare, the important thing is that we may in principle find all kinds of intermediate cases between ancestors who have only one living descendant and ancestors who have thousands of descendants. Further, *the more descendants an ancestor has, the more DNA matches will probably be found who has this ancestor in their family tree.* In other words, if a certain ancestor appears often in the family trees of your atDNA matches, this may in principle be due simply to this ancestor *having many descendants*. A corollary to this is that you cannot conclude that a certain historical person belongs to your ancestors just because you find him or her in the family trees of many of your atDNA matches. An alternative explanation is simply that this person has many living descendants that have tested their atDNA.

This raises the possibility that one reason why Jon Pehrsson (*1795) was found in so *many* family trees of the author's atDNA matches was that he had many now living descendants who had tested their atDNA. In other words, the *number of DNA matches* who had Jon Pehrsson in their family trees cannot be taken as evidence that he is one of the author's ancestors. What counts is the number of *segment triangulations* that were found, and especially the strong segment triangulations on chromosome 16, in combination with demographic data about his being a close neighbour to Carl Johan Forssén's (1828) mother in Pengsjö at the time when she got pregnant.

4.2. A remarkable coincidence?

As to the failure of the genealogical junctions analysis to identify the FFFFF/FFFFM couple Jon Jonsson (*1781) and Magdalena Eriksdotter Vänman (*1786) this could not be due to their having few descendants. As described in the results section they did not have fewer grandchildren (62 grandchildren) than the other ancestors at a similar genealogical distance. In this case, however, it was difficult to ignore another remarkable coincidence that was likely to nourish the imagination: The author's FFFFF Jon Jonsson (*1781) was a farmer at Pengsjö 3 during the same period that the author's FMMFM Greta Stina Olofsdotter (*1800) worked as a maid at Pengsjö 1 and the hypothesized FMMFF Jon Pehrsson (*1795) worked as a farmer at Pengsjö 2. In other words, during the same

time in the 1820s the author's FFFFF/FFFFM as well as his FMMFF and FMMFM all resided in this small village with only five farms. Was this simply a random coincidence?

Some other possibilities easily came to mind. First: could it be that Jon Jonsson (*1781) was the real father of Carl Johan Forssén (*1828)? Like Jon Pehrsson (*1795) he also seemed to be at the right place at the right time. However, a testing of this alternative hypothesis by means of segment triangulation (in the same way as had been done with Jon Pehrsson) did not provide any support for this possibility.

Another question that had to be raised was if Jon Pehrsson (*1795), rather than Jon Jonsson (*1781), could possibly have been the father also of the author's FFFF? Some evidence, however, clearly spoke against this possibility. First, Jon Pehrsson did not move to Pengsjö until 1823 or 1824, whereas the author's FFFF was born already in 1819. Secondly, if Jon Pehrsson (*1795) had been the father of the author's FFFF, he would belong to the same Y-DNA haplo group as the author (R-YP4123); some evidence, however, indicated that Jon Pehrsson (*1795) belonged to another haplo group (Q-BZ4901), which (if true) excluded the possibility that he could be the author' ancestor in the direct paternal line. The information on Jon Pehrsson's haplo group, however, was based on the Y-DNA testing of only one of Jon Pehrsson's descendants as reported on *Geni.com* and should therefore be treated with caution.

A third possibility was that Jon Pehrsson (*1795) and Jon Jonsson (*1781) might be close relatives, and that this explained why they chose to live as neighbours in the 1820s. A comparison of their family trees, however, did not show any genealogical connection between their families until six further generations back in time, in the 16th century. To summarize, the author's FFFFF and FMMFF being neighbours in the small village of Pengsjö in the 1820s might very well be just a remarkable coincidence.

4.3. Limitations and directions for future research

The analysis of genealogical junctions among unplaced DNA matches turned out to be a successful method in the two cases described in the present study. The usefulness of the method, however, needs to be tested also in other studies. For example, it may be of interest to explore different ways of defining the set of unplaced DNA matches that are to be included in the analysis.

In the present study, a cut-off of 100 cM was used for identifying the set of DNA matches, but this was a cut-off for the *combined* DNA shared by the two brothers with their atDNA matches. As can be seen in Table 1, only two of the 16 unplaced DNA matches (*Cesar-128* and *Bianca-102;* both descendants of the hypothesized FMMFF Jon Pehrsson) passed the cut-off of 100 cM for Brother 1. In contrast, six other unplaced DNA matches (*Sofia-150, Maja-132, Willy-124, Igor-124, Clara-114, and Axel-112;* all being related to the Norberg family) passed the cut-off for Brother 2. In other words: if only one of the brothers had tested their DNA the method would not have worked equally well. Furthermore, different sets of unplaced DNA matches would have resulted, depending on which brother did the testing. This suggests that the analysis of genealogical junctions may work best when at least two siblings are used to define the set of unplaced DNA matches.

Another important issue for further research concerns the confounding factor of the *number* of descendants of a given ancestor. Ideally, it would be a good thing if this confounder could be controlled for in some way when evaluating the strength of a given genealogical junction. For example, it might be that a genealogical junction which involves only two close DNA matches should be considered equally important as a genealogical junction

which involves four close DNA matches if the number of descendants for the couple in the former genealogical junction is only half the number of descendants for the couple in the latter genealogical junction.

Again, it should be noted that the cut-off of 100 cM in the present study was quite arbitrary. The possibly most important thing is to set this cut-off so that it produces a sufficient number of unplaced DNA matches. In the present study the cut-off of 100 cM produced nineteen unplaced DNA matches. In some cases it may probably more relevant to set a higher cut-off, whereas in other cases it may probably be more relevant to set a lower cut-off, depending on the number of unplaced DNA matches that are available.

Finally, more research should be done on segment triangulation to clarify the potentials and limitations of this methodology. Just as Bartlett (2016) argued for the importance of identifying *intermediate* MRCAs among DNA matches who share a given DNA segment when searching for the evidence of establishing ancestry, it is possible that there is potential for a similar search for more *remote* MRCAs beyond an established ancestor to explore that ancestor's ancestry. For example, Figures 4-9 do not only illustrate triangulations with DNA matches who had Jon Pehrsson (*1795) in their family trees but also triangulations with DNA matches who had Jon Pehrsson's parents, maternal grandparents, and great maternal grandparents as their MRCAs, thereby adding evidence not only that these DNA segments came from Jon Pehrsson (*1795), but also possible evidence of where he had had received these DNA segments from.

Further, Figure 6 suggests that it *might* be possible to trace DNA segments even further back in time, at least in some cases. When a systematic search was made for DNA matches who shared more than 15 cM of the specific segment on chromosome 1 (see section 3.5.3, pages 30-32 above), this did not only lead to the identification of four matches who had Jon Pehrsson (*1795) in their pedigree, one additional match who had Jon Pehrsson's parents in his pedigree, yet two matches who had Jon's maternal grandparents (MM/MF) in their pedigree, and two additional ones who had Jon's mother's grandparents (MMM/MMF) in their pedigrees. Quite surprisingly, it also led to the finding of yet another 32 DNA matches who had Jon Pehrsson's MMMM/MMMF Märeta Mattsdotter (*1678) and Lars Olofsson (*1675) in their pedigrees. Also quite surprisingly, several of these overlappings of DNA were quite large; twelve of these matches shared even more than 30 cM of the DNA segment.

What is to make of this kind of finding, where the number of DNA matches with a more *remote* MRCA sharing a given DNA segment *increases* in this way (as compared with the number of matches with less remote MRCAs)? A possible explanation would be in terms of pedigree collapse. Such a hypothesis was suggested by the fact that three additional DNA matches (i.e., who did not seem to have the Märeta/Lars couple in their pedigree), and who shared 37 cM, 31 cM, and 30 cM respectively of the segment, had a closer MRCA: the author's FMMFMMF Hans Jonsson Tiger (*1742) and FMMFMMM Magdalena Eriksdotter (around *1742). The father of Hans Jonsson Tiger is described in the genealogical sources as unknown, as is also the parents of his wife. This suggests the hypothesis that the Märeta/Lars couple might have a place somewhere in Hans' and/or Magdalena's pedigrees. If so, this would also mean that the Märeta/Lars couple would have more than one place in the author's pedigree. Maybe a pedigree collapse of this kind could help to explain the multiplication of DNA matches with the Märeta/Lars couple as MRCA who shared DNA segments on chromosome 1. It is a question for further research if it is even possible to answer such questions with segment triangulation methodology. It remains to be seen how far back in time this kind of procedure can be realistically pursued in a way that may result in trustworthy knowledge about ancestry.

References

Bartlett, J. (2016). CA and MRCA. http://segmentology.org/2016/01/02/ca-and-mrca/ : accessed 9 January 2023.

Bettinger, B. T. (2016) *The Family Tree Guide to DNA Testing and Genetic Genealogy.* Family Tree Books.

Bettinger, B.T. (2022). The shared cM project. Version 4.0. Retrieved from https://thegeneticgenealogist.com/2020/03/27/version-4-0-march-2020-update-to-the-shared-cm-project/

Eckeryd, R. (2017). *Acceptans eller utstötning? Ogifta mödrar i Ångermanland 1860–1940.* (PhD dissertation, Umeå university). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-132988

Thomas, T. (2021). Autosomal DNA and genealogy: How can DNA and triangulation be used to identify, evaluate and present conclusions of relatedness? *Journal of Genetic Genealogy, 9*, 1-171. Reference Number: 91.005

Tillmar, A., Fagerholm, S. A., Staaf, J., Sjölund, P., & Ansell, R. (2021). Getting the conclusive lead with investigative genetic genealogy – A successful case study of a 16 year old double murder in Sweden. *Forensic Science International: Genetics, 53*, 102525. https://doi.org/10.1016/j.fsigen.2021.102525

Journal of Genetic Genealogy

OUR STUDY: THE O'BRIEN IRISH ROYAL LINE, R-DC782, THE "Y" MALE LINE OF DESCENDANTS

By Dennis John O'Brien

Abstract

How the discovery of the Y-STR signature, "A Set of Distinctive Marker Values Defines a Y-STR Signature for Gaelic Dalcassian Families ", Dennis M Wright, JoGG Vol 5 Number 1 (Spring 2009) (a), led to the discovery of the R-L226 DNA Haplogroup, which lead to the discovery of a Haplogroup for a King and his descendants.

This paper builds on the discovery of this Y-DNA signature to establish that the SNP R-DC782, which dates from 900CE and is 7 splits after R-L226, the known Dalcassian Clan SNP, dated from 250CE (g), belongs to those who are descendants of King Brian Bóramha, AKA known as Brian Boru.

Thus, supporting a paper trail history over 32 generations in the Inchiquin Line leading to the current Leader of the Clan, Sir Conor Myles John O'Brien, 18th Baron of Inchiquin, as well as finding the missing descendant line from the senior royal line through Daniel O'Bryan, 1st Viscount Clare (c1577-c1663) (c).

As will be discussed later, there is evidence that those who carry this DC782 SNP are primarily of the surname O'Brien, while some have changed names because of political, religious reasons, adoptions, or because of a Non-Paternal-Event (NPE), with most having a known historical connection to the old Provinces of Munster and Thomond.

There are 94 people who have been tested through FamilyTree DNA (b) and carry the SNP R-DC782, 74 are members of the O'Brien Surname Project and carry a surname associated with O'Brien. There are 4 with the surname O'Dea, who presumed to be of the O'Brien line from associations and family swapping in the 14-15th Centuries.

That only leaves 12 other males of various surnames with no connections to the O'Brien's that can be confirmed at this time.

Introduction

History of Surname Project:

The O'Brien Surname Project (d) was created and announced on October 21, 2004 and registered with FTDNA (b). The first administrator was Michael O'Brien, of California, was a member of the O'Brien Clan Advisory Council.

One of the first members tested for the then 12 marker Y DNA, was Conor O'Brien, Clan Chief, and holder of the title "The O'Brien", the traditional Chief title going back a thousand years within the Clan. Conor is the 32nd descendant of King Brian Bóramha (Brian Boru) 942-1014CE as evidenced by paper trail and accepted family records.

There are many people with the name O'Brien, or one of it accepted variations (O'Bryan, Bryan, Brien, Bryant, O'Bryant, as examples)

In the Irish tradition the designation of "O" represents that the person is the "grandson" of Brian or Brien. In the period of 900CE to 1100CE there were several Clans where the name "Brian or Brien" was common for Clan Leaders. As the slow spread of surnames developed it was normal for a resident family to take of the name of their "Lord" or on whose lands they resided. Thus, the Surname O'Brien (O'Brian) became common.

However, within the Dalcassian Clans there was only one group who took on this Surname and that was those who were descendants of King Brian Bóramha (Brian Boru) or lived on the land holding of one of his descendants. The descendants of Brian Boru continued for many generations to be Provincial Kings or Princes and over time held extensive land holding in the Munster and Thomond Areas of Western Ireland.

Therefore, there are any people with the surname of O'Brien, but only a few who are DNA descendants of Brian Boru.

The O'Brien Surname Project was established and open to any person with the name O'Brien, or variation, based on a false concept that all such persons were related or had connection back to Brian Boru.

In the early days of "Y" testing, at 12 markers it appeared that the connection was very common and broad. It was only when the "Y" 25 and later the 37 markers became measurable, supported by the discovery of the R-L226 SNP in late 2009, that it became obvious that not all those of the name were in fact related. In some cases, non-Irish origins started to appear, especially those connected with Vikings, Normans and others who became integrated into the Ireland of the Middle Ages.

It was during 2008, that Michael O'Brien stepped down as Administrator and I was asked to take on that position from then, a position I still hold.

By December 2009 there were 150 members registered in the project. By this time the SNP R-L226 (Irish Type III) had been identified and was now recognised as the defining SNP for those who were descendants of the Dalcassian Clans. The O'Brien's were mainly that grouping.

As of January 2023, the project has 975 members, of which 254 have been tested to the Big Y level. In total 172 members have a related O'Brien Surname connection and are known or presumed to be carriers of the R-L226 SNP.

The balance of the O'Brien related surname members have been identified as belonging to other Haplogroups, mostly "R" related but also of the "B", "E", "G", "I", "J", "N", "Q" and "T" groups.
There are 63 members who though being R-L226 SNP, have no O'Brien related surname connection. Their surnames do relate, in most cases, to other known Dalcassian Clans and their results show SNP's that precede those of the R-DC782 haplogroup.

Methods and Data

Discussion
It is proposed that the haplogroup called R-DC782, a derivative of the R-L226 and the earlier R-M269 haplogroups, can be used to support that a male who test positive for R-DC782, is a descendant of King Brian Bóramha.

The paper by Dennis M Wright (2009) and the subsequent discovery of haplogroup R-L226 with clearly defined and exclusive makers in the DNA sequencing clearly established that those males with his SNP in the Ancestral Path are descended from members of the Irish Clan Dal gCais. This is further supported by the surnames of members of the L226 FTDNA Project which overwhelmingly are names associated with ancient and known sub-clans of this Greater Regional Clan.

If this conclusion is accepted and given that the surname O'Brien is one of the major sub-clan groupings, then it is

reasonable to expect that with the development of the Big Y testing regime we may expect that there would be a point in time when the male SNP results split between those of other sub-clans and the O'Brien's descending from the High King in the middle of the 900CE period.

The L226 Project shows hundreds of members who have the SNP L226 in their ancestral DNA history, of which 91 have tested positive to R-DC782. There are several other members who show indications that they would also test positive for DC782, however at this time those confirming tests have not been undertaking. Therefore, for this paper only those members confirm as DC782 in either the L226 or the O'Brien Surname Projects are being used for the analysis. Currently over 95% of those with positive tests appear in both projects.

After the discovery of the L226 SNP, it was the finding of the R-DC1 SNP in 2009 showing a significant split on the membership of the L226 project, supported by an almost exclusive use of an O'Brien related surname, that we realised that it became apparent there was a DNA path which exclusively cover the O'Brien sub-clan within the Dalcassian Clan.

This was further supported by the fact that Member 29355, Conor O'Brien was found to have a SNP of FGC13418. This SNP was also shared by member 477041, Tasman O'Brien. Both these person were related, being 4th cousins with a common ancestor of Sir Edward O'Brien, 4th Baronet of Inchiquin, (1773 – 1837), a 27th great grandson of Brian Boru.

The DNA connection (MRCA) proved that FGC13418 and an earlier SNP of FT120209 were common to both these members and they also were derived from the SNP YFS231286, believed to date from around 1200CE. (Refer to Chart-1)

Also, there are currently 32 members of the Projects who are also tested positive for this SNP. Within that group are a family of members named "Brien" who can show descendance through the senior line of the O'Brien's as descendant of Viscount Clare line(R-DC344) and thus connecting the Two separate, distantly related families of O'Brien back to the late 1400's CE at the time of the split in the Royal Irish line and the renouncement of the title of "King" in exchange for titles Earl of Thomond and Baron of Inchiquin.

Also, another group of O'Brien's who test positive for haplogroup, R-YSF231286, split after this SNP with DC1344, at around the early 1400's CE period, just before the historic changes to the titles of the senior O'Brien in late 1400CE. It is within the group that the family name O'Dea appears and would indicate that a minor O'Brien Family was use in securing the swap of sons for security or loyalty purposes. (*This is not part of this paper and may be subject to another paper in the future*.)

The one haplogroup that all these diverse yet related family members have is that they all test positive for R-DC782. This Haplogroup has been determining by FTDNA to originated in the 900CE timeline.

Therefore, it is not unreasonable to believe that Brian Boru either was the person whom R-DC782 SNP was originated or one of his sons.

While we know Brian had several brothers, all who died because of war or clan fighting and that they were originally from the "mac Cinnéide" Clan within the Dal gCais Clan, which eventually became known as the "Kennedy" surname family.

The Kennedy's who have been tested using the Big Y have shown negative for R-DC782 but positive for R-ZZ34_1 SNP which the immediate split before DC782 and dates form around 800CE.

No person with a surname associated with a non-O'Brien related surname, but still with a Dal gCais clan surname has tested positive to R-DC782.

Charts displaying various confirmation of haplogroup history and membership

Chart-1: **Ancestral Path of the Haplogroups from R-FGC5626 to R-L226 and then its SNP's sub-clads**
Displaying the changes in SNPs over the time period and the splits that occurred before and after R-DC782

Ancestral Path of the Haplogroups from R-FGC5626, to R-L226 and then its SNP's sub-clads

Examples of the spilts and time line between one, the O'Brien's and two the O'Brien's and Kennedy Sub-Clans

FTDNA Members 29355 & 477041 (4th cousins of the Inchiquin Line) - R-FGC13418

Steps	Haplogroup	Age Estimate	Time Passed	Immediate Descendants	Tested Modern Descendants
-1	R-FGC5626	750 BCE	1,350 years	2	800
0	R-L226	250 CE	1,000 years	2	799 (a)
1	R-FGC5660	300 CE	<100 years	2	629
2	R-Z17669	450 CE	150 years	5	599
3	R-ZZ31_1	500 CE	<100 years	4	400
4	R-FGC5628	550 CE	<100 years	3	321
5	R-FGC5623	600 CE	<100 years	3	253
6	R-FGC5659	650 CE	<100 years	4	228
7	R-ZZ34_1	800 CE	150 years	8	183
8	**R-DC782**	**900 CE**	**100 years**	**2**	**94 (b)**
9	R-Y5610	1050 CE	150 years	8	69
10	R-DC1	1150 CE	100 years	5	52
11	R-YFS231286	1200 CE	<100 years	2	32
12	R-FT120209	1400 CE	200 years	4	11
13	**R-FGC13418**	1800 CE	400 years	2	2

FTDNA Members N142662 & S112632 & IN73613 (descendants of the Viscount Clare Line) - R-FTB68795

11	R-YFS231286	1200 CE	<100 years	2	32
12	R-DC344	1450 CE	<100 years	3	7
13	R-DC310	1850 CE	400 years	3	5
14	**R-FTB68795**	1900 CE	<100 years	3	3

FTDNA Members 405921 & 422354 (Descendants of the Kennedy Lines) - R-DC951

7	R-ZZ34_1	800 CE	150 years	8	184
8	R-DC709	1050 CE	250 years	5	23 (c)
9	R-DC1544	1150 CE	100 years	4	7
10	R-DC951	1850 CE	700 years	2	2

(a) R-L226 is the defining haplogoup for the Dal gCais Clan

(b) R-DC782 is the proposed haplogroup for the decendnts of King Brian Boru - (95% recorded with this SNP have an O'Brien surname connection)

(c) Members testing negitave for R-DC782, this Kennedy line is an example of no O'Brien surname connection. This split occurs prior to the time lime of Brian Boru and falls in line with the seperation between the sub-clan of O'Brien's and the Kennedy's.

Chart-2: **Defined SNP Sub-clades for R-DC782**
Displaying the pathways of the DNA for the "senior" line of the O'Brien's

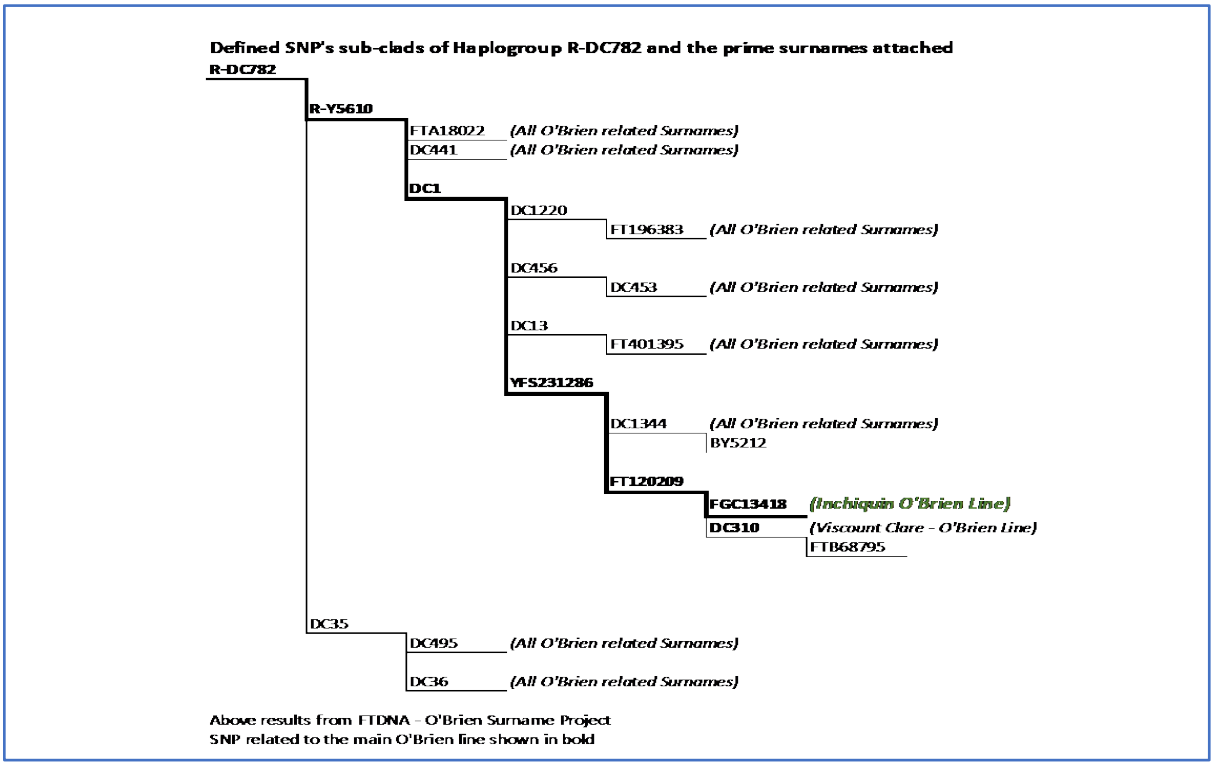
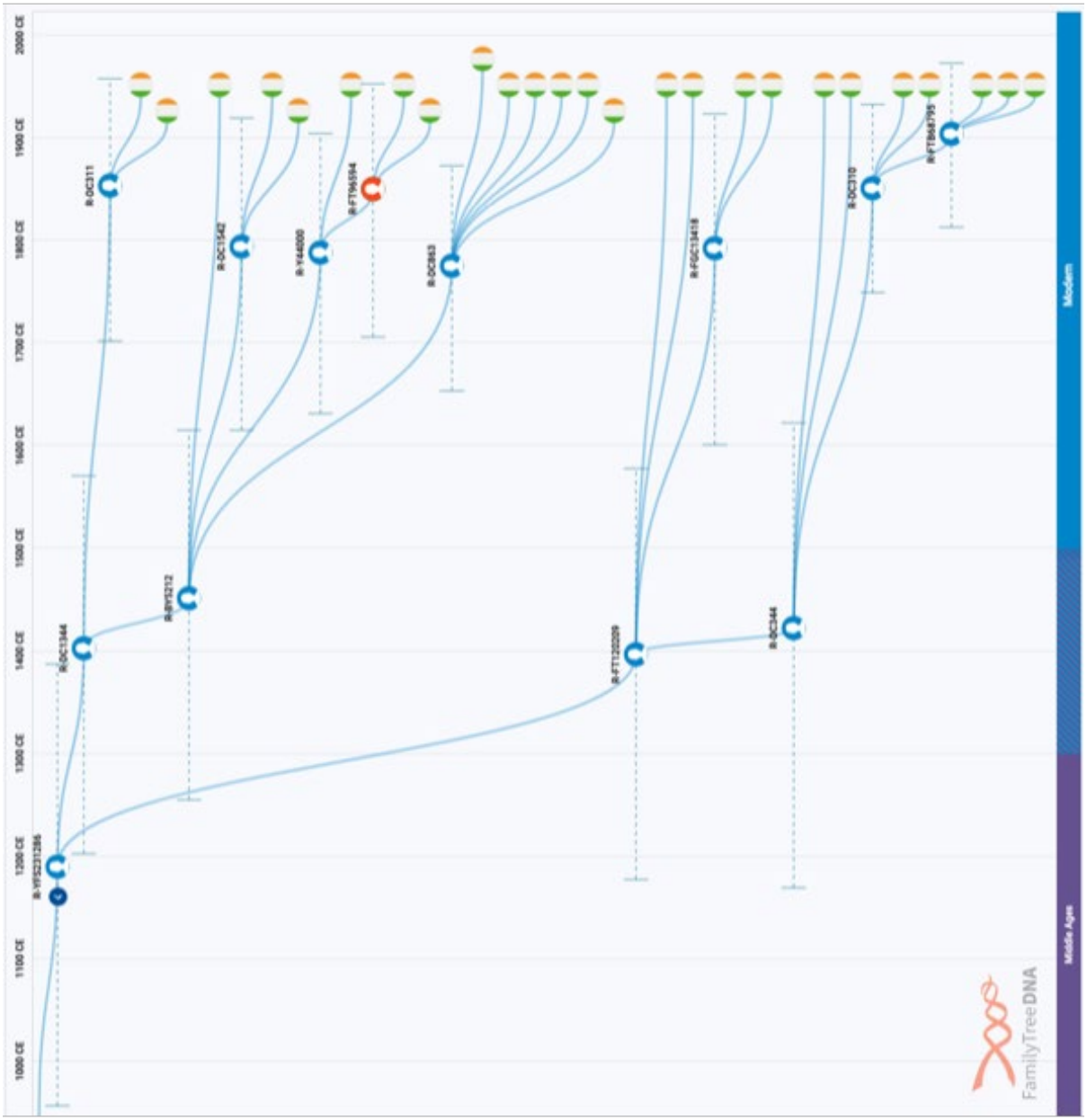


Chart-3: **Tree Time for Haplogroup R-YFS231286**



Conclusion

Based on the results of Y-DNA testing and the identification of a Haplogroup R-L226, which has been shown to belong to descendants of a small yet prominent Clan of Western Ireland in the Middle Ages, the Dal gCais Clan, which is well documented as the Clan from which the historical High King of Ireland, Brian Bóramha, was a member and head of this Clan, was a members, we can conclude that Brian would have been the carrier of the R-L226 SNP.

Unfortunately, his remains have been lost in time and the remains of his sons are not in places where access and therefore DNA samples can be obtained, we are then required to use logic as a bases for determining which downward SNP Brian would have carried or may have originated.

Given that the overwhelming results show that those with R-DC782 are of the surname O'Brien, which is derived from being a descendant of a "Brian". And given the continued blood and DNA lines in the Inchiquin and Viscount Clare families from the Princes and Kings of Munster and Thomond O'Brien family lines. And given that males tested for R-ZZ34-1, the precursor SNP to R-DC782, but negative for R-DC782, are not of the O'Brien Surname, but related to names associated with his earlier generations.

In addition, timelines for the haplogroup R-DC782, as determined through the records of FTDNA align with the period of Brian Bóramha known lifetime - 942-1014CE.

The DNA evidence for this O'Brien subgroup supports the paper trail lineage (mostly unusual among royal or noble documented lineages!) and that then further, assuming that the lineage is correct would mean that R-DC782 is the most likely haplogroup corresponding to Brian Boru himself.

The writer believes it is reasonable to claim that Brian Bóramha, born into the Dal gCais Clan, later to be crowned the High King of All Ireland and King of Munster and Thomond, is the progenitor of the haplogroup R-DC782 and that his descendant are the only carriers of this SNP.

Acknowledgements

Research support: Neville J Brien, Dennis M Wright

Endorsed by: Sir Conor Myles John O'Brien, Head of the O'Brien Clan of the Dal gCais Clan, 18 Baron of Inchiquin, 32nd descendant of King Brian Bóramha (Brian Boru) 942-1014CE, High King of Ireland and King of Munster and Thomond

Conflicts of Interest

The author declares no conflicts of interest and no commercial interests in the subjects covered by this study. He has no financial or personal interest in Family Tree DNA.

References

a) "A Set of Distinctive Marker Values Defines a Y-STR Signature for Gaelic Dalcassian Families", Dennis M Wright, JOGG, Vol 5 Number 1 (Spring 2009). https://jogg.info/wp-content/uploads/2021/09/51.002.pdf

b) FamilyTreeDNA, based in Houston, Texas USA, Now owned by MYDNA, Inc. a Melbourne, Australia. Website: https://www.familytreedna.com/

c) The "Brien's of Paramatta", Sydney Australia Website: https://www.parrabriens.com/

d) O'Brien Surname Project (FTDNA) Website: https://www.familytreedna.com/groups/obrien/about/background

e) Charts 1-3 have been produced using Surname Project information available to FTDNA Administrators.

f) In this text that following abbreviations have been used
   - DNA (deoxyribonucleic acid)
   - SNP (single nucleotide polymorphism)
   - STR (short tandem repeat)
   - yDNA or Y-DNA (y-chromosomal DNA)
   - FTDNA (Family Tree DNA)

g) While FTDNA says R-L226 originated ~250CE, YFull is suggesting ~500CE – (these differences relate to the size of the sampling base, which is larger in FTDNA.

General Background References

- The Irish Genealogist: the O'Briens of Dough and Ennistymon. By P. I. D. O'Brien and Kenneth Nicholls.
- ODavorens of Cahermacnaughten, Burren, Co. Clare. By George U. MacNamara.
- Kevin O'Brien's research on his DC344 line.

Journal of Genetic Genealogy

Quantum Genetic Genealogy Applications: A First Look

By Wesley Johnston (24 May 2023)

Overview

I first began monitoring quantum computing about 1997 as part of my role as head of Chevron's Advanced Information-Based Modeling Network. In 1999, I brought in a then-leader in quantum computers for an all-day seminar. So, I have been monitoring quantum computing for a long time.

For a very long time, people have been saying we are 10-20 years away from realizing the benefits of quantum computing. But now in 2023, several years after IBM introduced Qiskit[1] (pronounced kiss-kit) which allows anyone to run programs on IBM's quantum computers, productive applications are no longer 10-20 years away. They are here. And non-technical media such as "Time" magazine devotes cover stories to that reality.[2]

I created a web page[3] on which I present realistic ideas of what quantum computing can do for genetic genealogy that has not been possible before. I want to move beyond the pie-in-the-sky fantasy hype of popular coverage of quantum computing and present a realistic understanding of what quantum computing can do and what classical computers will continue to do since not everything is best done on a quantum computer.

I also want to put quantum genetic genealogy on the map as a very real and relevant topic for the future of genetic genealogy – because it has the potential to be a very significant component of the future of genetic genealogy and possibly a future not so far off as some think. Quantum computing is no longer “going to happen”. It has now been happening for several years. Theory is good but only when the focus is applying that theory in real applications. We need a consortium of those serious about quantum genetic genealogy applications to come together, share ideas and experiences, collaborate and develop applications.

I also want to ease the path for those thinking about the possibility of doing quantum genetic genealogy applications. I want them to know that they can use the resources that have been available now for several years and continue to grow rapidly. And I want to provide insights and tips to foster that active participation in quantum computing.

Some of what may appear to be hype, really is true: quantum computers CAN solve in a very short time what classical computers cannot ever solve because there are too many possibilities that would take eons for a classical computer to solve. But there are still things that classical computers do very well and will continue to do.

No "Killer App" Yet Identified

At this point in my investigation of quantum computing, I do not see a killer genetic genealogy app. What is there in genetic genealogy that is currently an NP-hard problem? Such a problem could be a killer app. But at this point, I do not see one. That probably is not because there is no such NP-hard problem and is more likely to be due to my limited level of knowledge of quantum computing that has not yet brought me to a point where I could see such an app – sort of like climbing a mountain and not being able to see something off in the distance until I am up high enough. So, I will keep climbing. But I would also like to see some discussion among genetic genealogists of what kind of NP-hard problems there might be that would be a good fit for a quantum computing program.

This Document

Part 1 gives the overview of what, based on my current level of knowledge, I see quantum computing being able to do for genetic genealogy and what classical computing will still do.

Part 2 has lessons learned from my own efforts to learn and do quantum computing – lessons from which others who want to do this (I hope there are more genetic genealogists who do want to do this) can learn so that their path is not as bumpy as mine.

Part 1 – Quantum Genetic Genealogy Applications

What Quantum Computing Does Best and What Classical Computing Will Still Do

Quantum Computing

- Quantum computers can deal with huge state spaces -- a great many possible combinations or possibilities -- which classical computers cannot deal with other than via sampling or probabilistic techniques that do not really deal with all the data.

- Quantum computers can handle large amounts of data, such as for analyzing and comparing large numbers of complex DNA sequence sets from testers. Whole genome sequences of large numbers of testers would no longer be computationally unmanageable.

- Quantum computing can help identify patterns in large genetic datasets that are not easily recognizable by classical computers. Note that this applies not just to genetic genealogy but to genealogy and history in cases where large data sets may hold patterns.

- Quantum computing can create better models of reality that fully take into account details that classical computers cannot include.

- Quantum computing might -- not really sure about this -- greatly improve phylogenetic analysis and tree-building.

- Doing a one-to-many comparison with every kit in a database would be something better for a quantum computer, possibly obviating the need for batch pre-processing and creation of "short kits" (I forget what the actual term is) such as GEDmatch does.

- CURRENT LIMITATION: For now, there is no programming language that elevates quantum computing to the level of modern programming languages. Quantum programming operates at the circuit level in specific values of qubits and actions of logic gates. In this sense, it is like the very early days of classical programming when you were really working at the level of machine language. So, there is a significant difference of conceptual levels that modern programming languages make unnecessary for working with classical computers than the machine-level operations you work with

in quantum programming.

- CURRENT LIMITATION: Both quantum qubits and classical bits can produce a result of only 2^n states. The difference is that the classical computer can only represent those states one at a time while the quantum computer can represent all of them at the same time through superposition, thus speeding up processing of combinations. But there is still a limit based on the number of qubits. The quantum computers currently available from IBM with Qiskit range from 5 to 14 qubits which can model 32 to 16,384 states. Thus only limited power of quantum computing is possible in such small quantum computers.

- CURRENT LIMITATION: The following is from IBM's "Quantum Computing in a Nutshell" web page: "As with the noise cancellation example above,[4] the amplitude and phase of qubits are continuous degrees of freedom upon which operations can never be done exactly. These gates errors, along with noise from the environment in which a quantum computer resides, can conspire to ruin a computation if not accounted for in the compilation process, and may require additional mitigation procedures in order to obtain a high-fidelity output on present day quantum systems susceptible to noise. Qiskit is capable of taking into account a wide range of device calibration metrics ... in its compilation strategy, and can select an optimal set of qubits on which to run a given quantum circuit. In addition, Qiskit hosts a collection of noise mitigation techniques for extracting a faithful representation of a quantum circuits output."[5]

  Here is a related comment by Martin Vesely on Stack Exchange[6]: According to so-called *threshold theorem*, it is possible to get rid of errors in quantum computation with arbitrary precision. However, there is an assumption that you have enough qubits. / To illustrate the idea, you can encode one qubit $|q\rangle=\alpha|0\rangle+\beta|1\rangle$ with more qubits, for example $|q\rangle=\alpha|0000\rangle+\beta|1111\rangle$ and after calculation, based on majority rule, to decide about result. / ... To conclude, it is possible to reduce noise but we need more qubits. Increasing number of qubits would theoretically leads to quantum processor with no (or at least low level) noise."

Classical Computing

- Quantum programs run on classical computers that pass information to the quantum computer.

- Classical computers will still do tasks that do not deal with huge state spaces.

- Classical computers are error-free and not subject to the "quantum noise" problem that is a reality with quantum computers.

- Classical computers will still do the repetitious processing of tasks such as comparing two kits to see if they match. It might be that a quantum computer could also do this, but the size of the dataset -- even for whole genome sequencing -- is small enough for a one-to-one comparison that a classical computer is really all that is needed.

Part 2 – Getting Started in Quantum Computing

Getting Started in IBM's Qiskit

If you want to use IBM's Qiskit (pronounced kiss-kit), it is best to follow the Qiskit website instructions.[7]

The Qiskit environment is very dynamic, improving often. Even a 3-year-old video on YouTube is out of sync with the current state of Qiskit. IBM really needs to have a definitive video on the Qiskit website that they keep current. Instead, all the YouTube videos by Qiskiteers that were once state-of-the-art can now lead you into frustration for setup, although they are still good for programming tips.

Bottom line: forget the YouTube videos for initial setup and follow IBM's instructions on the website.

IBM offers three options. I opted for their cloud version since it will always use the current version, as opposed to setting up a local version on your computer in an environment that you then have to maintain to stay in sync with Qiskit.

Andrew Helwer of Microsoft Research did a very good 2018 "Quantum Computing for Computer Scientists" video class (duration 1:28).[8] He gave a solid Mathematical basis for quantum logic gates, which are really transformations via matrix products. He kept it to Real numbers so that his unit circle state space is a lot easier to handle intuitively than is the Bloch Sphere. He does not cover the Math behind every quantum gate, but you can see it on this Wikipedia page on Quantum Logic Gates.[9]

A very useful tip on how to navigate the different Qiskit quantum computers to choose where to run your job and also nice code for monitoring your job in the queue is in this YouTube video.[10]

All programming in the Qiskit Lab is via interactive Python. Essentially, you write python code to define the problem and method, including creating a quantum circuit to which you pass decision-making in the state space. This constitutes a job that you then send to one of the chosen quantum computers, from which you receive a reply in the form of a final reading of the qubits that you have taken.

IBM has multiple levels of education about quantum computing, and they can be a challenge to find via navigation from the Qiskit web site. So, here are the links to each one.

- Qiskit Getting Started - https://qiskit.org/documentation/getting_started.html#
- Introduction to Qiskit (tutorial) - https://qiskit.org/documentation/intro_tutorial1.html
- Qiskit Tutorials - https://qiskit.org/documentation/tutorials.html
- Qiskit Textbook - https://qiskit.org/textbook/preface.html
- IBM Quantum Composer User Guide - https://quantum-computing.ibm.com/composer/docs/iqx
- IBM Quantum "Learn quantum computing: a field guide" - https://quantum-computing.ibm.com/composer/docs/iqx/guide/

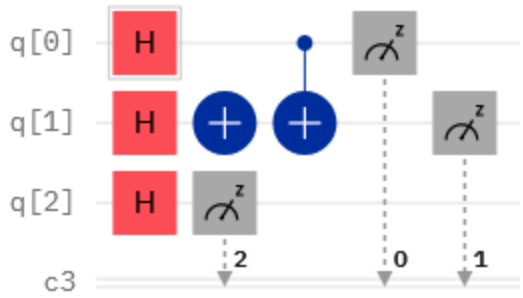
Toward Higher-Level Quantum Programming Languages

There is some progress out of the quantum paleolithic era when machine-level coding was the only way to go.

Microsoft Research's Azure Quantum[11] service (free for students but others pay) has a Quantum Developer's Kit which includes the Q# language[12] (Q-sharp, as in a musical note sharp, NOT Q-hash) language. You can see some Q# code in use on the 2018 Microsoft Research "Quantum Computing for Computer Scientists" video at about the 1 hour mark.[13] At least in that example, the main difference seems to be how quantum circuits are defined and referenced.

On a similar musical theme, IBM's Quantum Composer[14] lets you visually set up and simulate the use of quantum logic gates to build quantum circuits -- a very nice way to experiment with the different gates individually or in combination.[15] Keep in mind that simulations do not have the real-world quantum noise. But being able to interactively create quantum circuits of any configuration and instantly see how they behave is a great way to develop a more-intuitive sense of quantum circuit design.

When moving the gates into place, Composer also gives you, to the right, the modifiable code for the circuit that you have designed. For example, here is an arbitrary circuit slapped together as a visual example.



Note that the numbers associated with each measurement gate (the ones with the meter dial that point down to the c3 line) are not the output values but are the numbers of the qubits read by each measurement.

And here is the code on the right. The default is for 4 qubits, and I modified it to show and measure only 3 qubits, simply by changing the “[4]” default to “[3]” for both the qreg and the creg lines.

OpenQASM 2.0

Open in Quantum Lab

```
1   OPENQASM 2.0;
2   include "qelib1.inc";
3
4   qreg q[3];
5   creg c[3];
6   h q[0];
7   h q[1];
8   h q[2];
9   x q[1];
10  cx q[0], q[1];
11  measure q[2] -> c[2];
12  measure q[0] -> c[0];
13  measure q[1] -> c[1];
```

So, there are steps toward higher level languages, but essentially you are designing and running a quantum circuit in your program. So, it is very important to understand the quantum logic gates, which also means understanding quantum phase.

Quantum Gates and the Complex State Space

Some quantum computing logic gates are the same or similar to classical logic gates. This is especially true when dealing only with Real numbers. But even with Real numbers, some quantum computing gates have no analog in classical gates.

Quantum computing includes the full power of Complex numbers (the sum of a Real and an imaginary number). Imaginary numbers are those that have the square root of -1 as a factor. So, there are quantum gates (Y and S) that operate on qubits with imaginary numbers which greatly complicates how to understand what is going on in the quantum computer and how to visualize the Complex hyperspace in three-dimensional space.

I decided that it really is best for me to start out learning about quantum computing gates with Real numbers. But at some point you have to consider the Complex domain, the three dimensions we know and the fourth dimension of the imaginary component.

With Real numbers, we can use a unit circle to visualize the state space of the qubits of a quantum computer. But the inclusion of the imaginary dimension requires a unit sphere.

The main point of this section is that many quantum logic gates operate in ways that are not only not analogous to any classical logic gates but operate in a completely different realm. So, it is very useful to experiment with IBM's Quantum Composer to gain more understanding of what those gates do.[16]

Notation and the Search for Intuitive Understanding

There are two main hurdles to understanding quantum computing. The obvious one is the nature of quantum reality and most specifically quantum phases. The other one is the notation used to represent quantum operations on a quantum circuit.

Four main ways of representing the same thing, either in whole or in part, give different perspectives. Some offer more potential to having a "quantum epiphany" where it all (or mostly) becomes clear in an intuitive way – which is not easily achieved.

1. Circuit layout visualization

2. Linear algebra matrix representation
3. Bra-ket notation
4. Bloch sphere visualization

1-Circuit Layout Visualization

The visual representation of circuits is really like a new programming language, as noted above. Using IBM's Quantum Composer[17] gives the opportunity to experiment with different circuits, but by itself it really does not give understanding of quantum phases. And without understanding of quantum phases, one cannot truly understand quantum gates and quantum computing.

The simplest gate that has no corresponding classical logic gate is the H (for Hadamard) gate. The Hadamard gate puts a qubit into a state of quantum suspension, in which the qubit is neither 0 nor 1 but has a roughly equal probability of being read by a classical measurement as 0 or 1. The circuit, with the H gate followed by a measurement, looks like this.

The Hadamard gate can be considered as putting a coin into the state of being flipped in the air. That's a nice intuitive conception. The coin has not landed, so that it has an equal probability of landing as heads or tails. So, the output from the measurement of an initial value of 0 sent through a single Hadamard gate is 0 about half the time and 1 about half the time.

But, that really is not quite how the Hadamard gate works. So, when you put in a second Hadamard gate gate and read the output of this circuit

you might expect the result to be the same as two coin flips, still roughly half 0 and half 1. But what actually happens is that the second Hadamard gate takes the qubit back out of suspension to its original value of zero. And to understand this, the representation of the gate in matrix form makes it clear what is happening.

2-Linear Algebra Matrix Representation

Coming from a Mathematical background, I find the linear algebra matrix multiplication representation the most understandable form of representation of what a quantum circuit does, Although I have yet to master quantum phasing in even a basic way much less intuitively, the matrix multiplication representation is my "go to" form for understanding a quantum circuit's behavior.

A qubit value of zero can be considered as 2D vector with a 1 in the position for zero and a 0 in the position for one. And a value of 1 can be considered as a 2D vector with a 0 in the position for zero and a 1 in the position for one. Thus 0 and 1 are these vectors:

$$\begin{bmatrix}1\\0\end{bmatrix} \text{ and } \begin{bmatrix}0\\1\end{bmatrix}$$

The Hadamard gate puts a qubit into quantum suspension (the coin in the air that has not yet landed). It does this because as a matrix the Hadamard gate is

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}$$

What the quantum gate operations come down to is matrix multiplication in linear algebra.

So, applying H to a qubit in the zero state

means doing this matrix multiplication.

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$$

And applying two H gates to a qubit in the zero state

means multiplying the result of the first operation by the same Hadamard matrix.

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}1\\ 1\end{bmatrix}=\begin{bmatrix}1\\ 0\end{bmatrix}$$

Another way of thinking about it is that two Hadamard matrices applied to the same qubit are like multiplying the Hadamard matrix by itself, which yields the Identity Matrix.

Fundamentally, all the different notations and representations come down to matrix multiplications, although the elements of those matrices can be Complex and not as simple as with these Real number examples.

3-Bra-Ket Notation

Make no mistake about it: you ultimately need to think in terms of the bra-ket notation, which is the highest level and thus the simplest form – and thus has the most going on in a brief expression.

Bra-ket notation really is broader than its use in quantum computing. Quantum computing uses only the “ket” form (|0>) and not the “bra” form (<f|).

The ket notation (such as |0>) is a more concise version of the matrix multiplication representation. Ket notation is the way that those experienced with quantum computing give explanations. It has taken a while for me to look at something in bra-ket notation and understand just what it is actually saying. I still find myself having to convert the bra-ket notation to matrix notation so that I can fully understand the implications of the bra-ket representation.

4-Bloch Sphere Visualization

The Bloch Sphere (a unit sphere which is not unique to quantum computing) is a way of visualizing a 3D representation of the action of a quantum circuit on a qubit. The vertical axis is the imaginary axis. I found the unit circle in the Real number examples in Microsoft’s Andrew Helwer’s 2018 "Quantum Computing for Computer Scientists" video class (duration 1:28)[18] to be very understandable, but going to the sphere is a “quantum leap” that I find

challenging, particularly because quantum phases are not easily understood in this context for a beginner. I really found the Real number domain class very helpful for starting out with some sense of understanding. But quantum computing goes beyond the Real numbers so that it does require that "quantum leap" – that I have yet to achieve.

Symbols

In addition to representation, several symbols require understanding.

- |->
- |+>
- |ψ>

At this point, I just list these here to keep them on the map since I have to fully understand them

Quantum Phase

Understanding the quantum phase of a qubit is important to understanding how a circuit works as the signal passes through different gates.

I found the subject of quantum phase and how it relates to six of the quantum logic gates a very difficult step. The Field Guide "Quantum Phase" section[19] suddenly introduces an exponent of e (Euler's constant) that has two variables never previously mentioned and never fully explained. Suddenly, explanations are being given in terms of variables whose impact is not clearly defined. It seems that they are rotations of varying numbers of radians, but their impact is left completely unexplained so that developing an intuitive understanding from this section simply does not happen. The section makes for a poor explanation that requires extra experimentation and looking elsewhere for figuring out the full implications of what the section says and what the impacts of these six newly introduced gates are.

I find that the matrix multiplication conceptualization of the gates is easier to understand initially. The Wikipedia page with the matrices of all the quantum logic gates[20] helps a lot to understand them in terms of matrix multiplication.

Current Status

In the absence of a motivating application and the absence of a good text or vide to understand quantum phase (and how it impacts notation), I stopped my strong initial effort to come up to speed on quantum computing. I have done some minor – virtually trivial – looks now and then. But I really need a

dedicated block of time to find a way to come up to speed on quantum phases, the symbols and the notation so that I can progress further.

---

[1] https://qiskit.org/
[2] https://time.com/6249784/quantum-computing-revolution/
[3] https://wwjohnston.net/famhist/quantum-genetic-genealogy.htm
[4] Refers to how radio signals are made clearer by cancelling noise with a properly shifted sound wave
[5] https://qiskit.org/documentation/qc_intro.html
[6] https://quantumcomputing.stackexchange.com/questions/12148/exponential-growth-of-noise-in-quantum-computers
[7] https://qiskit.org/
[8] https://www.microsoft.com/en-us/research/video/quantum-computing-computer-scientists/
[9] https://en.wikipedia.org/wiki/Quantum_logic_gate
[10] https://www.youtube.com/watch?v=aPCZcv-5qfA
[11] https://learn.microsoft.com/en-us/azure/quantum/overview-azure-quantum
[12] https://learn.microsoft.com/en-us/azure/quantum/overview-what-is-qsharp-and-qdk
[13] https://www.microsoft.com/en-us/research/video/quantum-computing-computer-scientists
[14] https://quantum-computing.ibm.com/composer/docs/iqx
[15] https://quantum-computing.ibm.com/composer/docs/iqx
[16] https://quantum-computing.ibm.com/composer/docs/iqx
[17] https://quantum-computing.ibm.com/composer/docs/iqx
[18] https://www.microsoft.com/en-us/research/video/quantum-computing-computer-scientists/
[19] https://quantum-computing.ibm.com/composer/docs/iqx/guide/introducing-qubit-phase
[20] https://en.wikipedia.org/wiki/Quantum_logic_gate

Journal of Genetic Genealogy

Fully Calculating Autosomal DNA Coverage Recursively

By Wesley Johnston, May 2023

Abstract

In my first paper ("Calculating Autosomal DNA Match Coverage: A generalized additive recursive method"[i]), I gave the pseudo-code for a generalized additive recursive algorithm to accurately calculate the lower bound of the range of DNA coverage for a target person.[ii] In that paper, Appendix 4 ("Calculating the Shared DNA Component: Dealing with Ranges") provided an extended set of scenarios showing how the coverage of a person for whom none of their children tested can only accurately be reported as a range, with a lower bound and an upper bound. Like the cone of uncertainty in a hurricane forecast, there is no one single value that can be considered to be the coverage estimate when no children of the target person have tested.[iii]

Dr. David Stumpf made the first implementation of the algorithm in his Graphs for Genealogists (GFG) software. Jonny Perl then created the DNA Painter "Coverage Estimator" tool, which implemented the method of Paul Woodbury as used by Leah Perle Larkin, a method that propagates the average (instead of the lower bound) up the generations from the test takers to the target ancestor. So, the DNA Painter implementation and both the algorithm in my prior paper and in GFG will give different single numbers, neither of which fully represents the range that is the reality of the coverage for the referenced ancestor in no-child-tested scenarios.[iv]

In this paper, I extend the algorithm to accurately calculate both the lower and the upper bounds of the range. I also explore the behavior of the ranges and their average and of the propagated average (as used in DNA Painter). I then provide an intuitive understanding of DNA coverage ranges to understand who the best target tester is to improve the overall coverage of the target ancestor.

Review of the Recursive Algorithm

My lower-bound algorithm consisted of two parallel operations. One operation traced all descendants of the target person to see which ones had DNA-tested. Thus, this operation is essentially a top-down tracing. The second operation does a bottom-up calculation of the DNA coverage.[v]

1. Count the number of children (N) of the reference person (in the first call, this is the target ancestor).
2. Calculate the number of pieces P of the Venn diagram of combinations of the children (see Figure 1) and the maximum coverage percentage M for each. P=(2^N)-1, and M=1/(2^N).
3. Generate the module-internal tables, setting values to be calculated initially at zero.
4. For each child, determine the coverage of the child by calling this same module.
5. For each piece/row p of module-internal table, calculate W(p) and R(p) for that piece/row.
6. Add the percentages for the pieces together (sum the R(p)) to calculate the coverage percentage of the reference person and return that number and exit the module.

The key vehicle for all the calculations is the Venn diagram of all permutations of combinations of the children's DNA contribution to the parent's DNA coverage. The Venn diagram is what enables the recursive calls that can accurately calculate the lower and upper bounds.[vi]

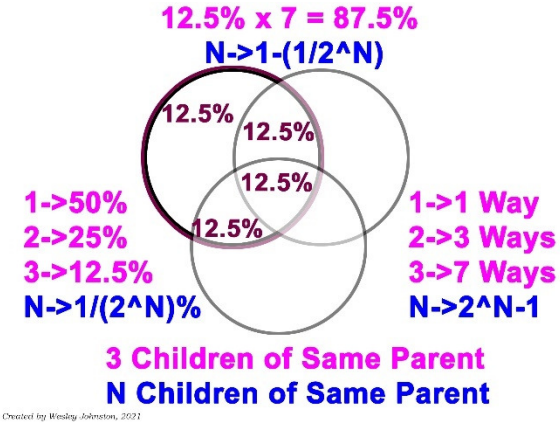


Figure 1 Pieces of DNA Covered by N Children (P = Ways with M% in each Piece)

The module-internal table represents the Venn diagram for the person. For a person with 2 children (thus a Venn diagram of two intersecting circles with one piece unique to each child and the third intersecting piece shared by both children), the table looks like this, where N is the number of children and P is the number of pieces of the Venn diagram so that P = (2^N) -1.

Lower Bound Module-Internal Table - People and Pieces With Completed W and R Values in Call Level 1

P (Piece)/N (child)	1	2	M	W(p)	R(p)
1	0	1	25%	50%	12.5%
2	1	0	25%	100%	25%
3	1	1	25%	100%	25%

Each row represents a piece of the Venn diagram.

There is a column for each child of the target person and the columns for M, W(p) and R(p). The child column bit value represents the presence (1) or absence (0) of that child in that piece of the Venn diagram. So, row 1 in the example table represents the piece of the Venn diagram that considers only the DNA contribution of child 2.

In this way, each piece of the Venn diagram represents a unique combination of the children: the table is functionally identical to the Venn diagram.

Differences in Calculation of the Lower and Upper Bounds

In this table, child 2's child has tested, and child 1 has tested. So, child 2 has 50% coverage, and child 1 has 100% coverage. The example shows the calculation of the lower bound for the parent of the two children. The lower and upper bounds each require their own Venn diagram and thus their own table to

accurately calculate the bounds of the range of DNA coverage of the target person from the DNA of their children. Column M is the same in both tables. But W and R are calculated differently.

Lower Bound (Max/Product)

In the lower bound case, in pieces of the Venn diagram for combinations of the children's DNA, the worst-case scenario (the lower bound) is the contribution of the child with the largest DNA coverage. This piece of the Venn diagram reflects DNA inherited from the parent by all the children in that piece. So, the parent's own coverage from that piece will be at least the amount due to the contribution of the child who has the most coverage.

Thus, for each row/piece, W (the combined weight of the children's contributions) is calculated simply by taking the coverage from the child with the maximum DNA of any child represented in that row/piece.

M = 1 / (2^N) expressed as a percent (the maximum coverage possible for a piece of the Venn diagram)

W(p) = Maximum of all products of each child's presence/absence bit times their own coverage, for row p

R(p) = M * W(p) for row p

Each weight, W(p), is calculated for its row by using the combination of children specified in the binary columns for that row and multiplying that 0 (absence) or 1 (presence) value by the DNA coverage of that child. The maximum value of each of these products then becomes the value of W(p) for that row p.

For example, in the table's row/piece 1, cell(1,1) is 0, and cell(1,2) is 1. So, W(1) is the greater of (0 * 100%) and (1 * 50%), which is the greater of 0% and 50%. Thus W(1) = 50%.

The Result (R) for that row is the product of the maximum (M) and the amount of coverage of the child with the highest contribution (W(p).

Upper Bound (Sum/Min)

In the upper bound case, in pieces of the Venn diagram for combinations of the children's DNA, the best-case scenario (the upper bound) is the aggregate contribution of all the child. While this piece of the Venn diagram reflects DNA inherited from the parent by all the children in that piece, if they have less than 100% themselves, then the upper bound assumes that the DNA that they inherited in this piece does not match the DNA the other children inherited in this piece. So, the parent's own coverage from this piece will be either (a) the sum of the M-weighted contributions of all the children or (b) the maximum amount (M) that this piece can cover of the parent's DNA, whichever is least.

M = 1 /(2^N) expressed as a percent (the maximum coverage possible for a piece of the Venn diagram)

W(p) = Sum of all products of each child's presence/absence bit times their own coverage times M, for row p

R(p) = Minimum of M and W(p) for row p

Using the same two children as in the lower bound table, the upper bound table looks like this.

Upper Bound Module-Internal Table - People and Pieces With Completed W and R Values in Call Level 1

P (Piece)/N (child)	1	2	M	W(p)	R(p)
1	0	1	25%	12.5%	12.5%
2	1	0	25%	25%	25%
3	1	1	25%	37.5%	25%

For example, in the table's row/piece 3, cell(3,1) is 1, and cell(3,2) is 1. So, W(1) is the sum of (1 * 100% * 25%) and (1 * 50% * 25%) = 25% + 12.5% = 37.5%.

The Result (R) for that row is the minimum of (a) the maximum (M) and (b) the sum (W(p)) = Min(25%, 37.5%) = 25%.

In this case, because one child has DNA-tested, the lower bound and upper bound are the same. In cases where two or more children of a parent have not DNA-tested but do have tested descendants, a single number cannot fully represent the DNA coverage. Only a range, defined by a lower and an upper bound, can fully represent the DNA coverage. In such cases, the lower bound will always be less than the upper bound.

Understanding How the Calculation of Coverage Works in the Recursive Algorithm

The first call of the algorithm considers the target person and his/her children. Each child is then checked to determine their own DNA coverage. If the child has DNA-tested, then their coverage is 100%. Otherwise, the exact same algorithm is called for the child and their children. And this process repeats until all DNA-tested descendants of the target person have been found and their contributions included in the calculation. Their lower and upper bound contributions propagate up through the call levels to allow calculation of the target person's DNA coverage estimate.

Considerations of Using DNA Coverage Estimates

It is very important to keep in mind just what these estimates are – with the main point being that they are estimates and thus uncertain. Actual DNA from actual relatives will, with high probability, yield an actual result somewhere within the bounds of the coverage range. But there are important things to keep in mind when dealing with theoretical situations as we are with any algorithmic estimation. I have put together several relevant points in Appendix A, "Estimates Based on Averages".

Full Coverage Excel Visual Basic Program

My prior paper provided only pseudo-code. And what is shown above provides the pseudo-code extension of the algorithm to full coverage of both the lower and upper bounds. However, in this paper, I want to understand the behavior of the bounds and the propagated average. So, I needed to write a program in some language to calculate both the lower and upper bounds for any configuration of a target ancestor and their DNA-tested descendants.

Here are the reasons that I decided to implement the program in Excel Visual Basic, using a GEDCOM file as input.

1. I wanted a program that anyone can use with their own GEDCOM file to calculate the upper and lower bounds of DNA coverage of anyone in their file who has DNA-tested descendants. In particular, I did not want to use a language that required downloading any software.
2. I wanted something that is completely transparent, with relatively simple code that is relatively easy to understand. I did not want to complicate things by using a language that starts array counting at zero instead of one.
3. What I would really like is for every family tree software product to implement calculation of DNA coverage as a standard feature. It is not enough for Legacy Family Tree to have webinars about DNA coverage. They really should implement it in their Legacy Family Tree software, as should every other family tree software. The Visual Basic program makes it extremely obvious how very simple it is to implement robust DNA coverage estimate calculation.

The full details of the Excel Visual Basic program, including the code itself, are in Appendix B. Appendix C presents a module by module overview of the program.

Behavior of the Bounds and Averages

While my prior paper showed that all but recent ancestors require ranges and not single numbers to accurately report their DNA coverage, I really wanted to run test cases to see how the bounds and averages behave in different scenarios – different configurations of a target ancestor and their DNA-tested descendants in the family tree structure.

The Mama Pepa Scenario

The primary motivation for my deep dive into DNA Coverage has been a project to reconstruct the DNA of Josefa Ruiz, known to her descendants as Mama Pepa. Many of her descendants have DNA-tested. So, I really wanted to see how well they cover Mama Pepa's DNA.

Here is the complete tree. The lowest level person(s) in each branch has done an autosomal DNA test. I have highlighted the test takers in bold underlined red.

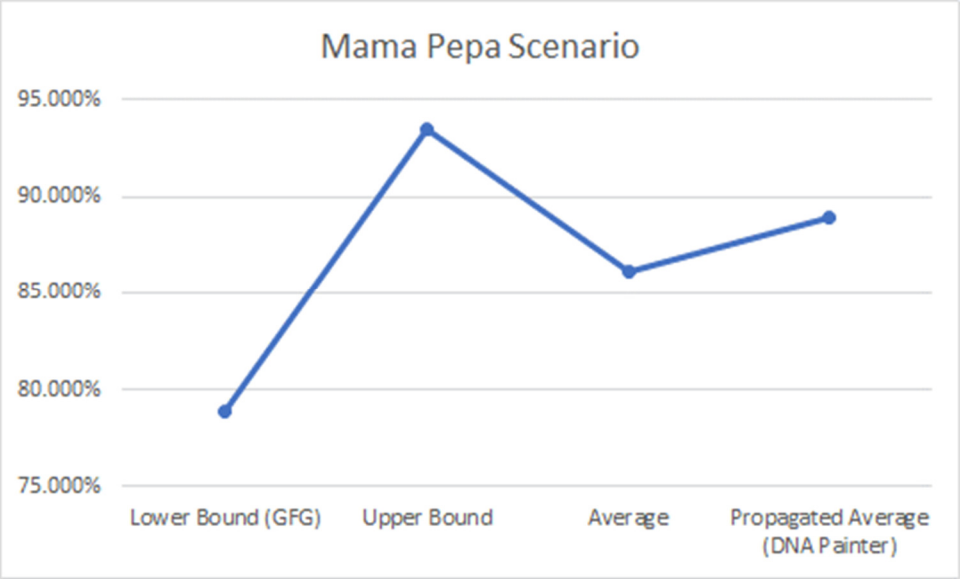
1-Josefa Ruiz
|-----2-Rita Salazar
| |-----3-RitaDau1
| | |-----4-RitaGrCh1
| | |-----**5-RitaGrGdChild1**
| |-----3-RitaDau2
| | |-----**4-RitaGdDau2**
| |-----3-RitaDau3
| |-----4-RitaGdDau3
| |-----**5-RitaGrGdDau3**
|-----2-Adalberta Salazar
| |-----3-AdalbertaCh1
| | |-----4-AdlabertaCh1Ch1
| | |-----4-AdalbertaCh1Ch2
| | |-----4-AdalbertaCh1Ch3
| | |-----5-AdalbertaCh1Ch3Ch1
| | |-----**6-AdalbertaCh1Ch3Ch1Ch1**
| |-----3-AdalbertCh2
| | |-----4-AdalbertaCh2Ch1
| | |-----**5-AdalbertaCh2Ch1Ch1**
| |-----3-AdalbertaCh3
| | |-----**4-AdalbertaCh3Ch1**
| |-----3-AdalbertaCh4
| |-----4-AdalbertaCh4Ch1
| |-----**5-AdalbertaCh4Ch1Ch1**
|-----2-Amalia Salazar

| |-----3-AmaliaCh1
| |-----4-AmaliaCh1Ch1
| |-----5-AmaliaCh1Ch1Ch1
|-----2-Mercedes Salazar
| |-----3-MercedesCh1
| | |-----4-MercedesCh1Ch1
| | | |-----5-MercedesCh1Ch1Ch1
| | | |-----5-MercedesCh1Ch1Ch2
| | |-----4-MercedesCh1Ch2
| | | |-----5-MercedesCh1Ch2Ch1
| | |-----4-MercedesCh1Ch3
| |-----3-MercedesCh2
| | |-----4-MercedesCh2Ch1
| | | |-----5-MercedesCh2Ch1Ch1
| | |-----4-MercedesCh2Ch2
| | |-----4-MercedesCh2Ch3
| |-----3-MercedesCh3
| | |-----4-MercedesCh3Ch1
| |-----3-MercedesCh4
| | |-----4-MercedesCh4Ch1
| |-----3-MercedesCh5
| |-----3-MercedesCh6
| |-----3-MercedesCh7
| | |-----4-MercedesCh7Ch1
| |-----3-MercedesCh8
| |-----4-MercedesCh8Ch1
|-----2-Maria Salazar
| |-----3-MariaCh1
| |-----3-MariaCh2
| | |-----4-MariaCh2Ch1
| | |-----5-MariaCh2Ch1Ch1
| |-----3-MariaCh3
| |-----3-MariaCh4
| |-----4-MariaCh4Ch1
| |-----4-MariaCh4Ch2
| |-----5-MariaCh4Ch2Ch1
|-----2-Manuel Salazar
|-----3-ManuelCh1 Salazar

The 27 DNA-tested descendants range from 3 to 6 generations of descent from Mama Pepa through six of her children, some of whom have sizable numbers of tested descendants so that their own DNA coverage is high.

Here are the results.

	Rita	Adalberta	Amalia	Mercedes	Maria	Manuel
Lower	34.375%	56.250%	12.500%	92.773%	84.375%	50.000%
Upper	50.000%	73.633%	12.500%	98.669%	85.938%	50.000%



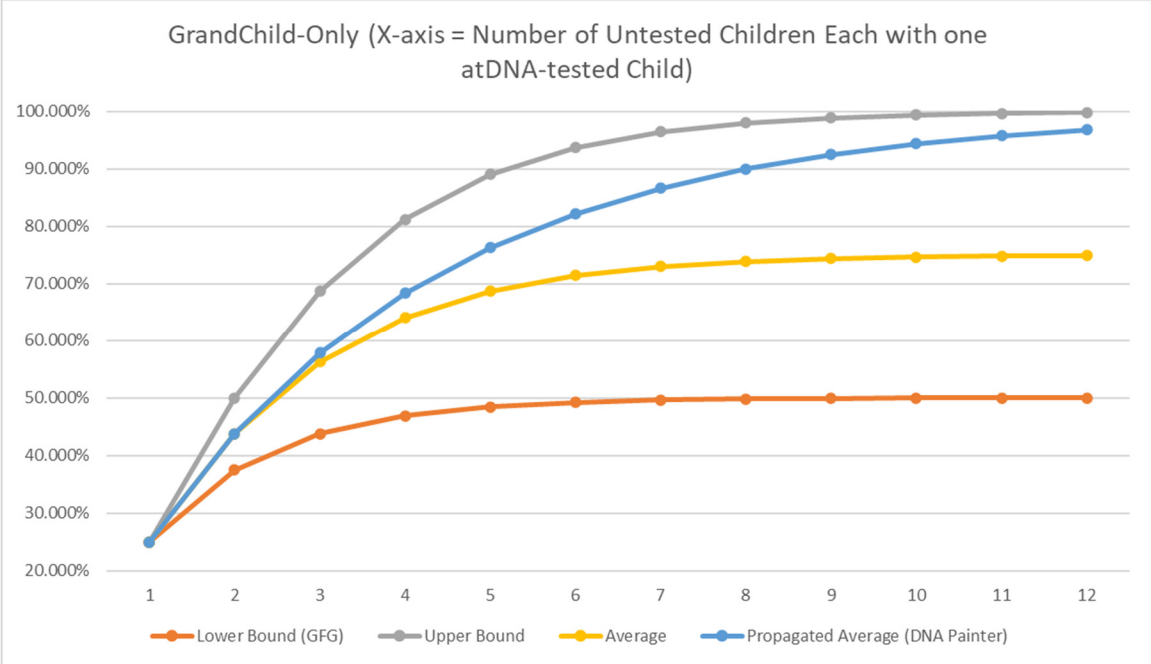
Lower Bound (GFG)	Upper Bound	Average	Propagated Average (DNA Painter)
78.906%	93.442%	86.174%	88.945%

This result aligns with what I expected prior to any of the testing. The lower bound and upper bound are relatively close, making a range of only 14.5%. And the average of the bounds is close to the propagated average. The closer bounds force convergence of the averages.

Even though the propagated average (implemented in DNA Painter) is a bit more optimistic than the average of the bounds, they really are not that far apart. And even though the lower bound (implemented in Graphs for Genealogists) is more pessimistic than the average, they differ only by 7.2%.

Grandchild-Only Scenario

In the first of two suites of stepwise tests, the target person has children who have not done a DNA test but who each have one child who has done a DNA test. Here are the results.



Tested Grand Children	Lower Bound (GFG)	Upper Bound	Average	Propagated Average (DNA Painter)
1	25.000%	25.000%	25.000%	25.000%
2	37.500%	50.000%	43.750%	43.750%

3	43.750%	68.750%	56.250%	57.813%
4	46.875%	81.250%	64.063%	68.359%
5	48.438%	89.063%	68.751%	76.270%
6	49.219%	93.750%	71.484%	82.202%
7	49.609%	96.484%	73.047%	86.652%
8	49.805%	98.047%	73.926%	89.989%
9	49.902%	98.926%	74.414%	92.492%
10	49.951%	99.414%	74.683%	94.369%
11	49.976%	99.683%	74.830%	95.776%
12	49.988%	99.829%	74.909%	96.832%

Consideration of the Bounds

I found this surprising, since I expected the lower bound and the upper bound to converge. Instead of converging, the lower bound asymptotically approaches 50% and the upper bound 100%. So, their average approaches 75%.

In hindsight, this makes perfect sense. The lower bound is the contribution of only one child in combinations while the upper bound is the aggregated contribution of all the children. None of the untested children of the target person will ever have more than 50% coverage. So, using just one of them to calculate the lower bound will never have more than 50% coverage for the lower bound of the target person.

Consideration of the Propagated Average and Lower Bound

The propagated average is overly optimistic (just as the lower bound is overly pessimistic). The propagated average always exceeds the average of the upper and lower bounds and thus increasingly converges on the upper bound while diverging from the lower bound and from the average of the bounds. So, the propagated average implicitly assumes that a best-case scenario is more likely than a worst-case scenario. The expectation, however, is that both the worst-case scenario and the best-case scenario are equally likely so that their average is the most likely overall scenario.

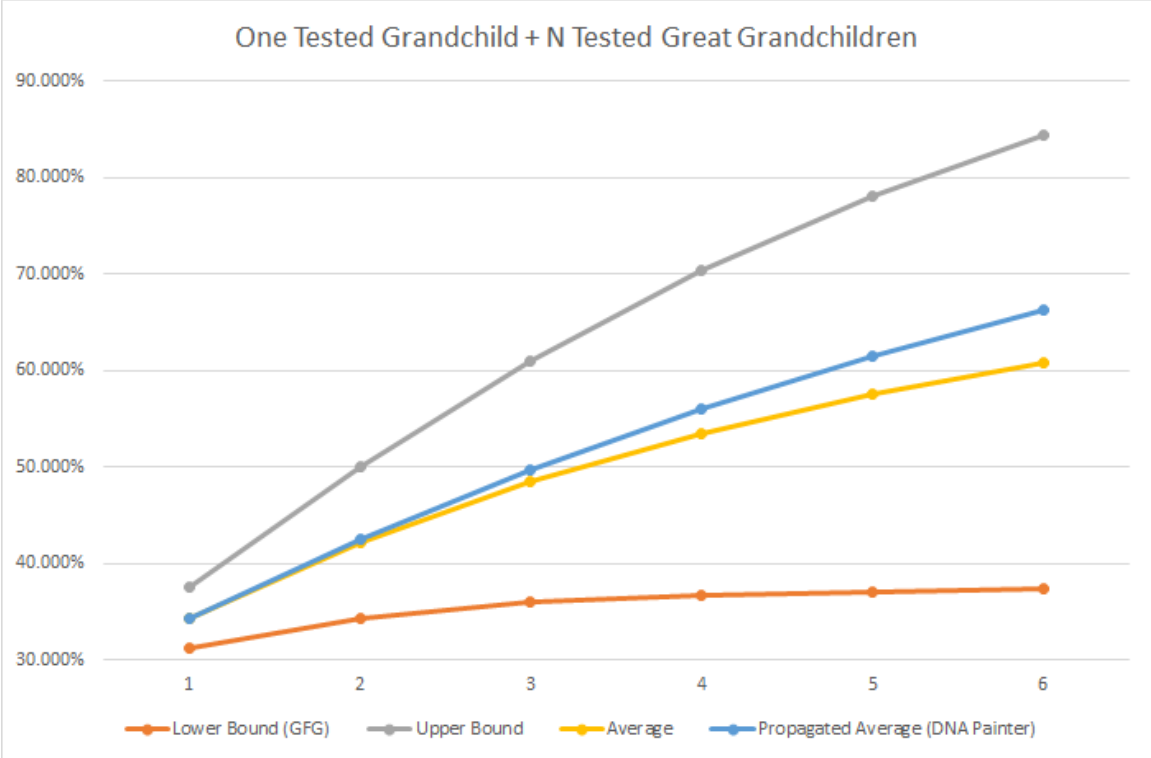
While the propagated average for the scenarios with four or fewer children is at or close to the average of the bounds, with 5 or more children the propagated average starts at 11% more than the average of the bounds and diverges from there so that by 12 children, the propagated average is 29.3% above the average of the bounds and only 3% below the upper bound.

Taking into account the Mama Pepa scenario and the grandchild-only scenario, the behavior of the propagated average and of the lower bound both prove most useful as single-number estimates when the target person has children who either have tested or whose own DNA coverage is high.

One Tested Grandchild + N Tested Great Grandchildren

In the second suite of stepwise tests, the target person has one untested child who has one tested child (the target person's grandchild) plus one or more untested children of the target person who each has one untested child who each has one tested child (a great grandchild of the target person).

Here are the results.



Tested Great Grand Children	Lower Bound (GFG)	Upper Bound	Average	Propagated Average (DNA Painter)
1	31.250%	37.500%	34.375%	34.375%
2	34.375%	50.000%	42.188%	42.578%
3	35.938%	60.938%	48.438%	49.756%
4	36.719%	70.313%	53.516%	56.036%
5	37.109%	78.125%	57.617%	61.532%
6	37.305%	84.375%	60.840%	66.340%

The lower bound asymptotically approaches 37.5%. This is as expected since the one tested grandchild provides 25% coverage of the target person in each case while the combined coverage from the tested grandchildren asymptotically approaches 12.5%,

The upper bound still approaches 100% since it is based on the aggregate of all the test takers.

The propagated average still skews to the optimistic side but does not diverge as rapidly from the average of the bounds since the one tested grandchild moderates it.

An Intuitive Understanding

The way that I find the most intuitive for thinking about the scenarios is to focus on the parent-child family group of the targeted ancestor as the parent. We know that one DNA-tested child covers 50% of the parent, two tested children cover 75%, and N tested children cover 100 – (100/(2^N)) percent of the parent's autosomal DNA.

The intuitive leap is that situations in which a child's own coverage approaches 100% make the situation approximate the child having done a DNA test.

The key is the impact of the child's coverage on the lower bound. When you can raise the lower bound so that it is closer to the upper bound, all the estimates are squeezed into a smaller range. And you can do that if more of the children are at or near 100% coverage themselves. **For targeted testing, the best new person is the one who will raise the lower bound the most.**

For configurations with only a few children of the target person, using the propagated average in DNA Painter's Coverage Estimator is good. If the number of children increases beyond 4 without the children having a good amount of coverage themselves, then the propagated average becomes overly optimistic. Similarly, the Graphs for Genealogists use of only the lower becomes overly pessimistic in the same scenarios. In such cases, you really do need to know the lower and upper bounds and their average.

But if you have children of the target person whose own coverage is high, then the lower bound increases so that all the estimates become closer to each other.

DNA Coverage as a Standard Feature of Family Tree Software

Autosomal DNA testing began in 2007. It is now 2023. More than twenty million people now have the taking of of an autosomal DNA test as an actual life event. Yet, none of the family tree software companies have a standard "Autosomal DNA Test" life event. It is long past time for the family tree software companies (and the GEDCOM standard) to have a standardized "Autosomal DNA Test" life event.

Now that the full calculation of autosomal DNA coverage estimation is shown to be easily calculated with a relatively simple program, the family tree software companies should provide standard reports of the lower and upper bound estimates of DNA coverage of anyone in the database.

Appendix A: Estimates Based on Averages

First and foremost, keep in mind that the calculated DNA coverage provides only an estimate and not an exact result. The output depends on the input. Only the raw DNA of the testers can provide exact results.

Several factors make these estimates in the right ballpark but not exact.

Average Inheritance

The estimates use the average amount that a person inherits from a specific parent: 50%. The Shared cM project[vii] by Blaine Bettinger (on the DNA Painter website) uses actual data from known relationships to see the actual range of values of shared centiMorgans for each relationship. While a child inherits on average 3,485 cM from a specific parent, the actual range is 2,376 to 3,720 cM. If we assume that the average is 50% (and thus that the total cMs of a person are 6,970), then the range is from 34% to 53%.

Testing Company Differences: 50% of What?

Briton Nicholson has focused on two more issues. One of those is the differences that the testing companies have for total number of cMs of a person.[viii] Here are the totals he cites.

Site	Half (cM)	Full (cM)
GEDmatch	3,587	7,174
23andMe	3,537	7,074
MyHeritage	3,500	7,000
AncestryDNA	3,475	6,950
FTDNA	3,384	6,768

The Shared cM Project average of 3,485 cM differs least from the AncestryDNA count. But it ranges from 51.5% (for Family Tree DNA) to 48.6% (for GEDmatch). So, even the Shared cM Project average is not average, depending on which company's results you use.

The algorithm used in this present document does not account for these testing company differences and uses 50%.

Gender Inheritance Differences

Briton Nicholson also pointed out that the transmission of autosomal DNA varies depending on the gender of the parent.[ix] Here is what he wrote about a simple model he made for autosomal DNA inheritance that deals with the same issue we face with estimating DNA coverage: "For this model, a simple assumption is made that men and women pass DNA to their children in the same manner. In reality, DNA from fathers recombines at lower rates than from mothers. So the variability in shared DNA could be greater from fathers, especially for successively all-male lines. Not taking into account the sex of the parent makes the model vastly simpler to implement and still provides useful approximations, especially for lines that are a fairly even mix of males and females."

The algorithm used in this present document does not account for these gender inheritance differences.

Practicalities of Ancestor DNA Reconstruction

Kevin Borland created Borland Genetics to go beyond theoretical calculations of coverage of an ancestor's DNA by descendants. With Borland Genetics, you can reconstruct an actual DNA kit of the ancestor's DNA from the DNA of the descendants. Kevin Borland in 2019 explained why actual reconstruction of a DNA kit of an ancestor will not result in as much DNA as the theoretical coverage calculations, especially in cases with unphased test result.[x]

Dr. David Stumpf has implemented "in-silico" reconstruction in Graphs for Genealogists of an ancestor's DNA from the actual results of tested descendant.[xi]

Overall Impact on DNA Coverage Estimates

The overall impact on DNA coverage estimates based on the 50% average of DNA inherited by a child from a specific parent is in the right ballpark but should not be taken as an exact percentage. Even the estimates of lower and upper bounds are not hard-line bounds since the Shared cM Project results show that actual cases exist with cMs less than and more than the 50% average.

Appendix B: Full Coverage Excel Visual Basic Program

What the Program Does

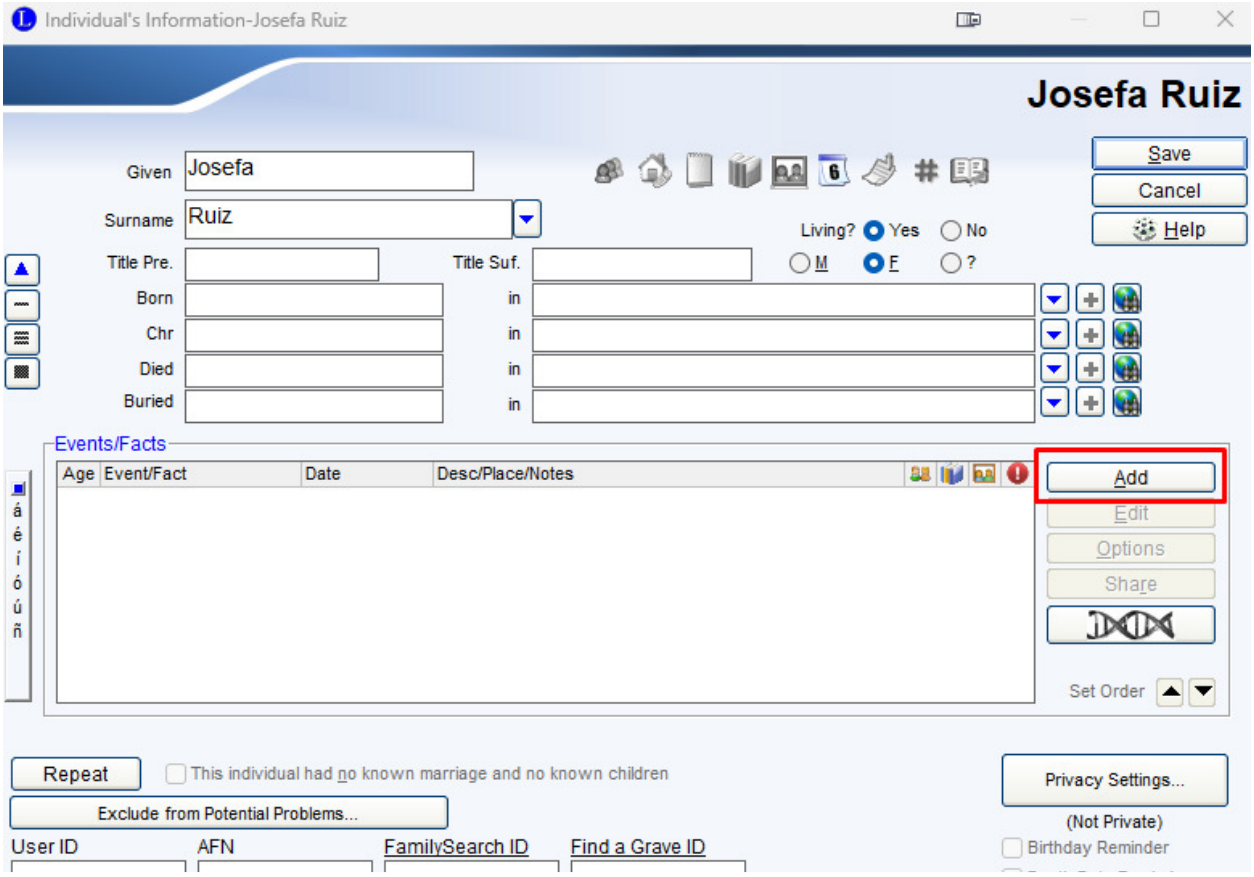
You provide the program with two inputs: a GEDCOM file and the RIN (Record Identification Number) of the target person in the GEDCOM file. The program then reads the GEDCOM file and then steps through the family tree starting with the target person and going one generation at a time to identify all DNA-tested descendants of the target person. The program then propagates the children's lower and upper bounds of DNA coverage back up one generation at a time until the target person's lower and upper bound DNA coverage can be calculated. At the end of processing, it populates the Excel worksheet "Results Report" with the complete report, generation by generation, of the bottom-up calculation of the DNA coverage of every person on the line between the target person and the test taker.

Preparing your GEDCOM file

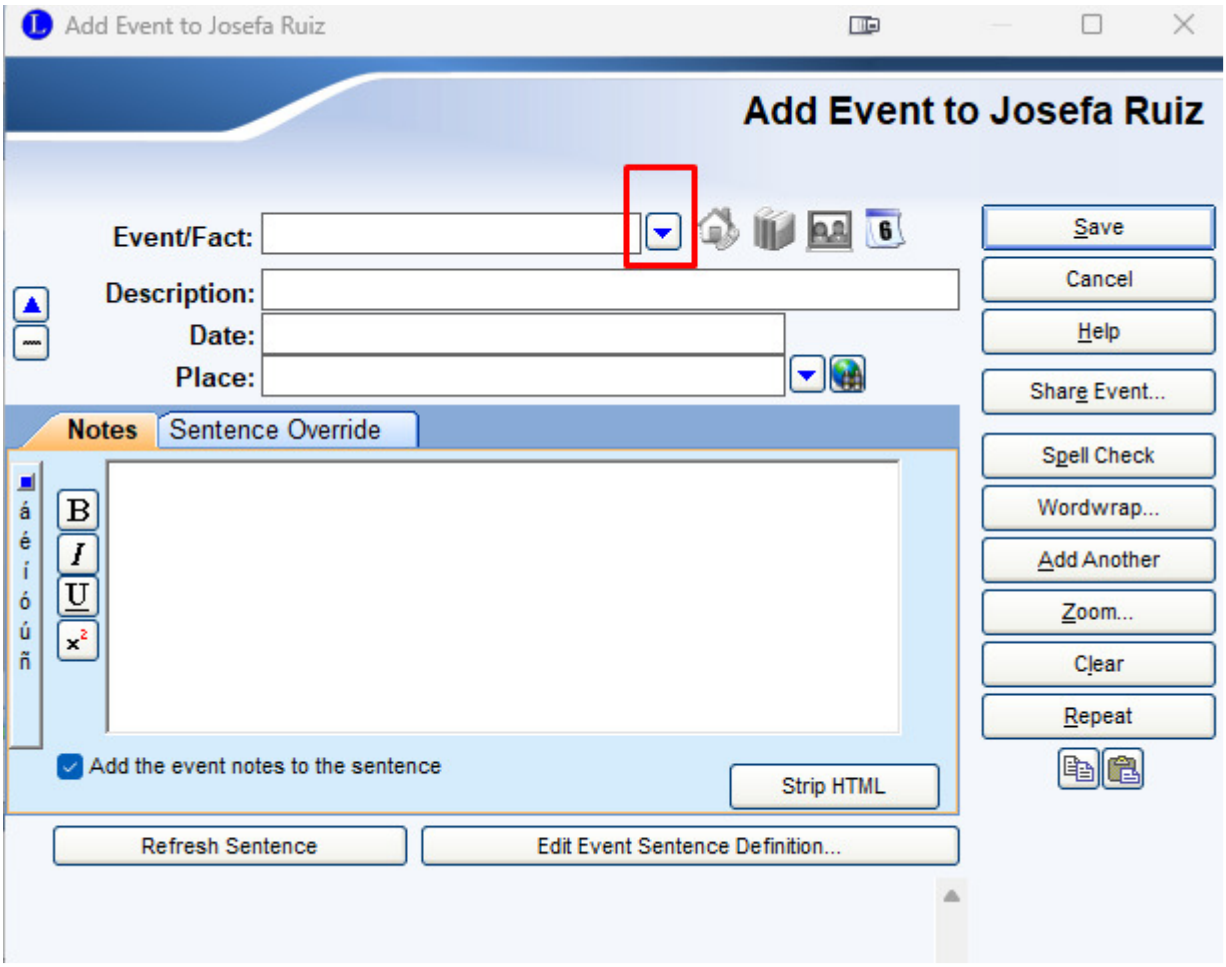
No family tree software company currently has an event to indicate that a person has taken an autosomal DNA test. Thus, you will have to use the family tree software in which you have your database so that you can create a user-defined event which the program will look for as "atDNA".

Here is how I have done this in Legacy Family Tree software.

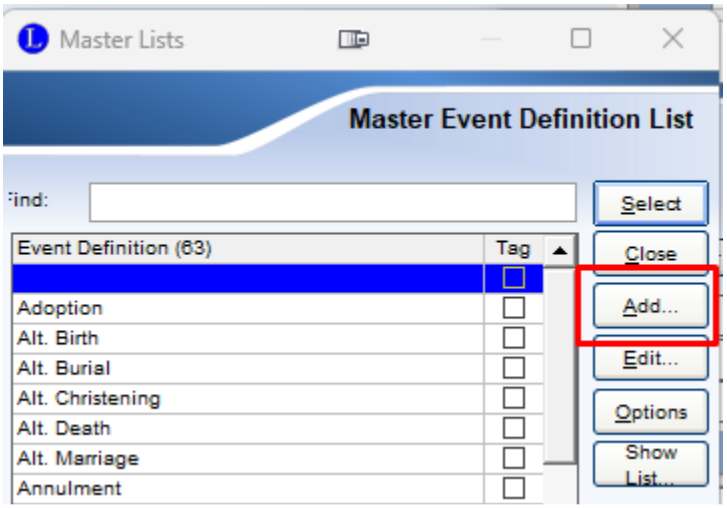
In the "individual's Information" window for a person, click on the "Add" button to add an event just as you would to add any other life event.



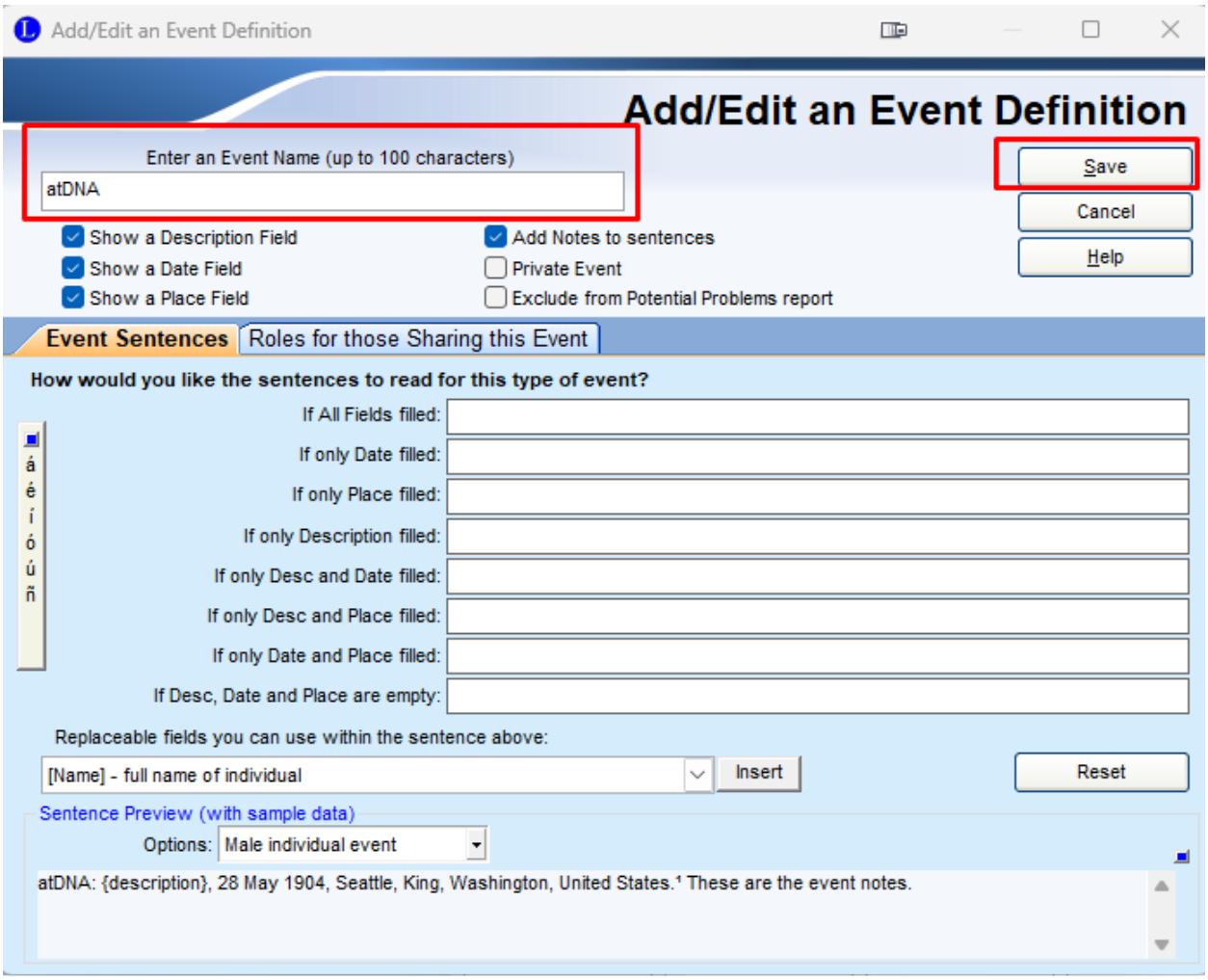
This opens the “Add Event” window where you need to click on the down arrow by the empty “Event/Fact” pane.



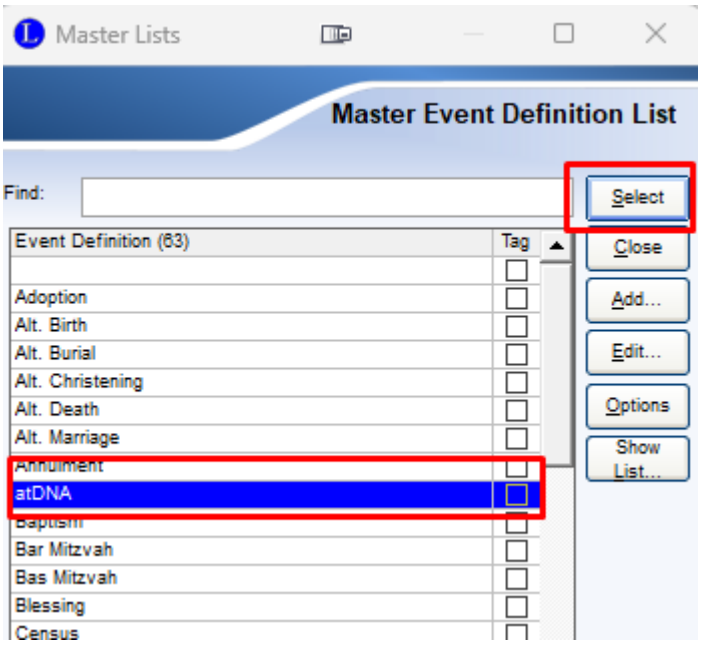
This opens the Master Event Definition List where you need to click on “Add”.



This opens the “Add/Edit an Event Definition” window where you type “atDNA” into the Event Name pane and then click Save.



Now when you go back to the Master Event Definition List, you will see “atDNA” as an event that you can click on and then click “Select”.



What this will do when you export your database to a GEDCOM is to create the following line in the GEDCOM for this person.

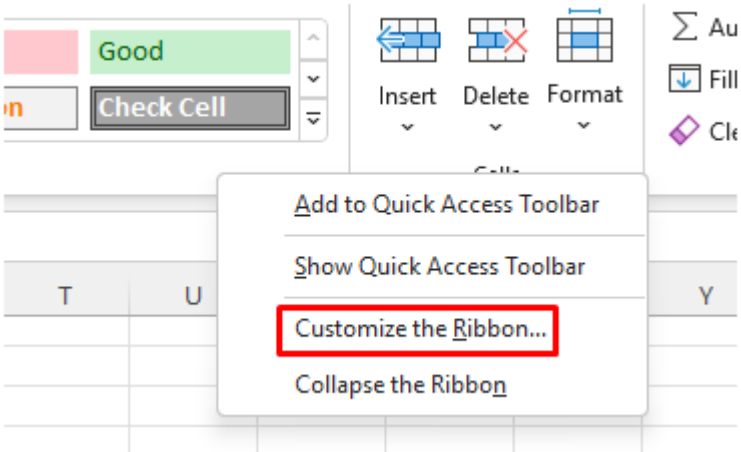
```
1 EVEN
2 TYPE atDNA
```

The Excel Visual Basic Program searches for the "2 TYPE atDNA" record in the GEDCOM file to identify which people in your database have done an autosomal DNA test. So, you need to enter the atDNA event for all those in your database who have done an autosomal DNA test.

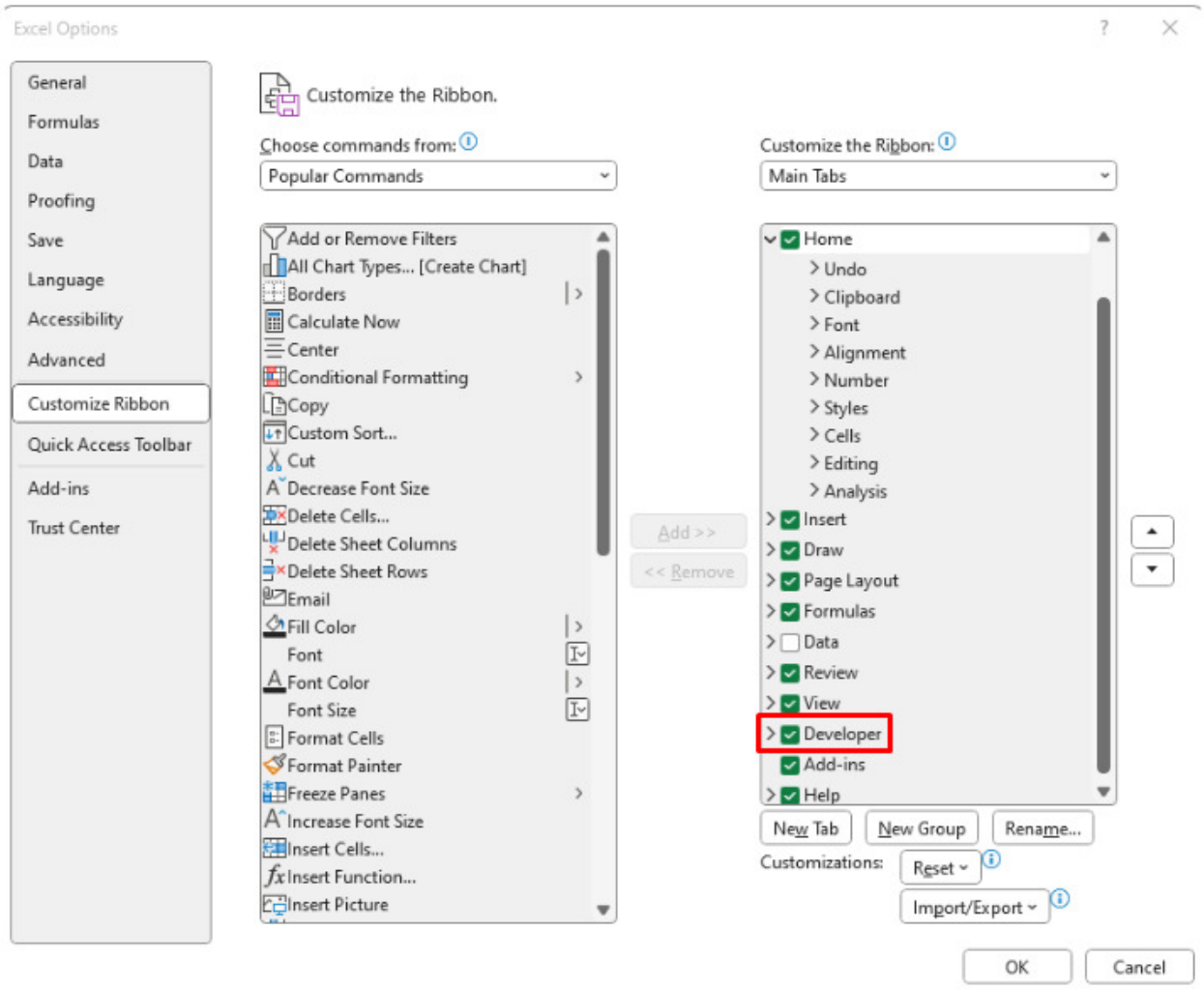
Enabling Visual Basic in Your Excel Spreadsheet

The program runs in Excel. However, you must first enable the Developer tab in Excel. Start with a blank Excel spreadsheet, and name the active worksheet "Results Report".

In Excel, right click on the ribbon at the top of the spreadsheet and click on "Customize the Ribbon…".



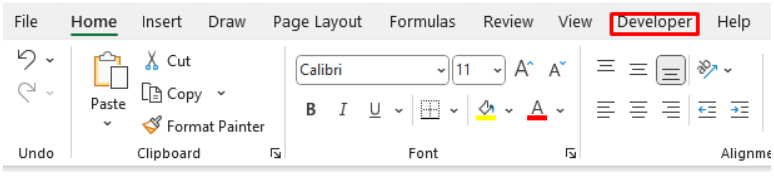
This will open the Excel Options window where you simply want to click on the check box to enable "Developer". Then click "OK".



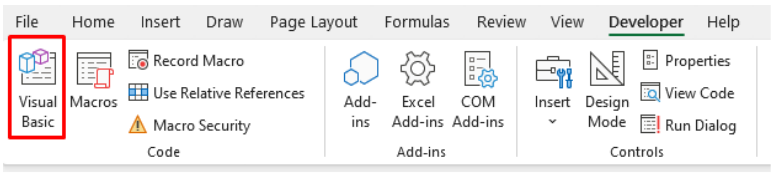
This completes the setup of the Developer tab and its functions for you in Excel.

Setting Up the Visual Basic Program in Excel

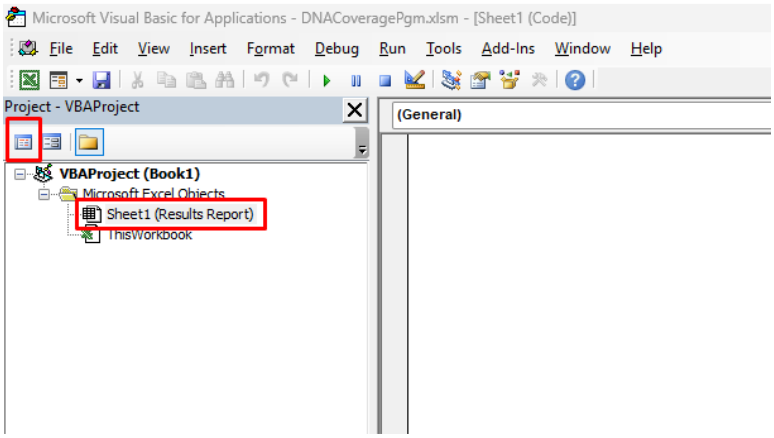
When you now see your ribbon at the top of your Excel worksheet, click on the "Developer" tab.



This will open the Developer tools ribbon where you click on "Visual Basic" to open the Visual Basic editor.



This will open a completely new window “Microsoft Visual Basic for Applications”. Since you already named the worksheet in your main window “Results Report”, this is what the new window looks like.



Click on “Sheet1 (Results Report)” and then on the leftmost icon above it. This will change the right-side pane from the gray background to what is shown in the above image.

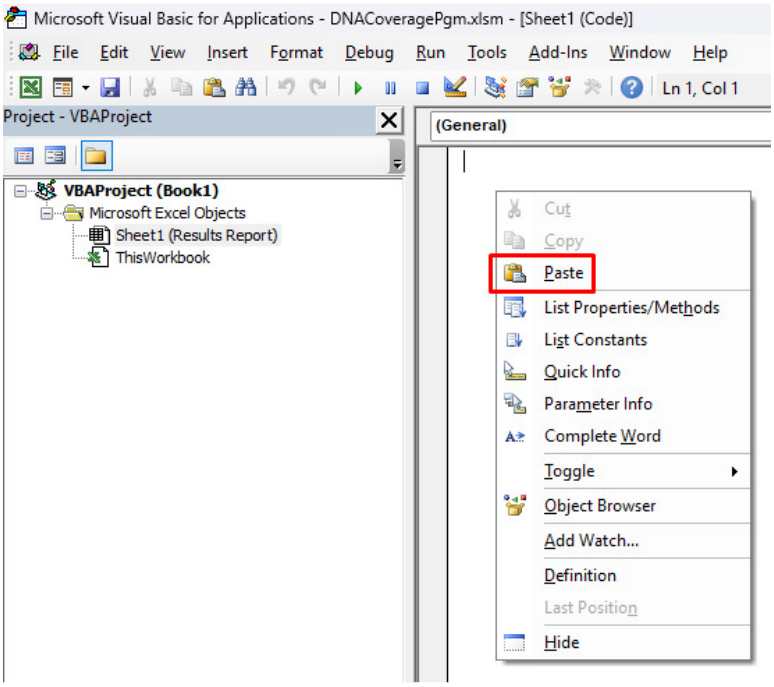
What you want to do now is to copy the complete text of the program from this paper and paste it into the right-side pane. Select all the text of the program code from this paper, starting with the line that reads

```
'*************************************************************
```

And ending with the line that reads

```
End Function ' CalcLDNACoverage
```

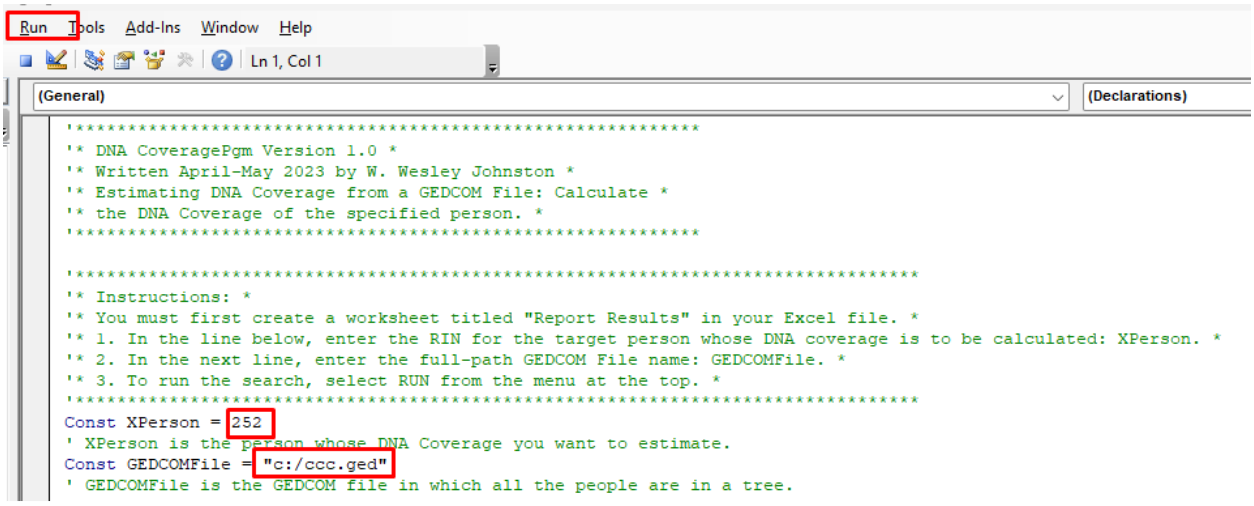
Then right click in the right-side pane of the Visual Basic window and choose the “Paste” option.



Then do a File/Save to save the spreadsheet with the code in it. Since this is now a Visual Basic spreadsheet, Windows will save it as filetype XLSM and not the usual XLSX. This way you have completed the setup of the Visual Basic Program.

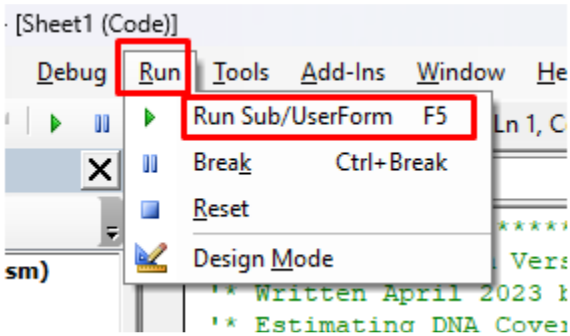
Running the Excel Visual Basic Program

You need to enter the RIN of the target person, as found in your GEDCOM file. And you need to enter the path that tells the program where to find your GEDCOM file. Both of these are near the top of the program code in two different lines.

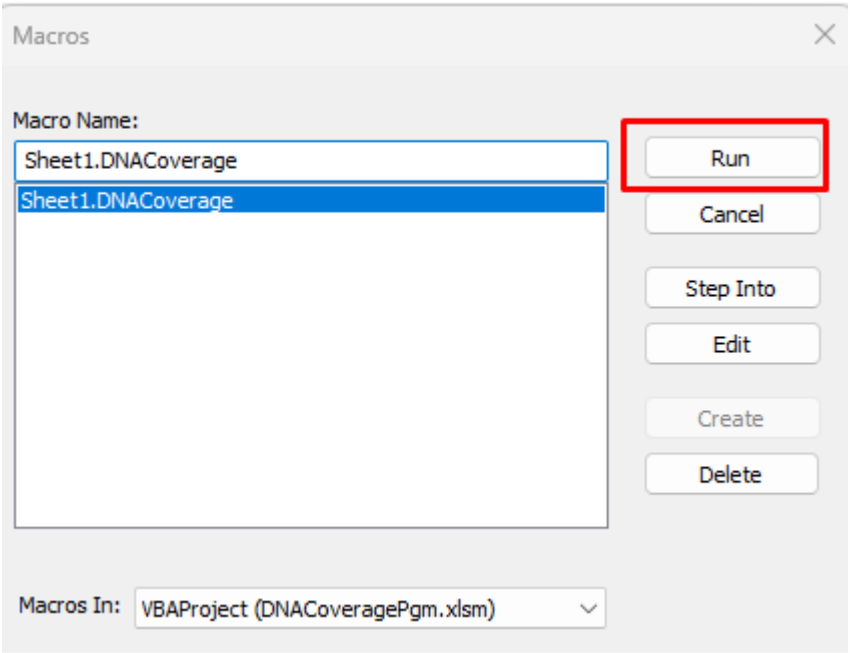


The default setup is to calculate the DNA coverage of the person with RIN 252 in the GEDCOM file that is located at c:/ccc.ged on your computer. So, these are the values you will have to change.

That is all you have to enter. Then go up to the ribbon at the top and click on “Run” and then on “Run Sub/Userform”.



This will open a window labeled “Macros” where you just click again on Run.



The program will run, popping up small windows as it completes milestones. When it says “Done”, the “Results Report” worksheet will display the complete steps in the calculation first of the lower bound and then of the upper bound.

Understanding the Excel Visual Basic Program

If you want to know what is going on “under the hood”, see Appendix C.

The Excel Visual Basic Program Code to Copy

```
'*************************************************************
'* DNA CoveragePgm Version 1.0 *
'* Written April-May 2023 by W. Wesley Johnston *
'* Estimating DNA Coverage from a GEDCOM File: Calculate *
```

```
'* the DNA Coverage of the specified person. *
'*********************************************************

'*************************************************************************************
'* Instructions: *
'* You must first create a worksheet titled "Report Results" in your Excel file. *
'* 1. In the line below, enter the RIN for the target person whose DNA coverage is to be calculated:
XPerson. *
'* 2. In the next line, enter the full-path GEDCOM File name: GEDCOMFile. *
'* 3. To run the search, select RUN from the menu at the top. *
'*************************************************************************************
Const XPerson = 252
' XPerson is the person whose DNA Coverage you want to estimate.
Const GEDCOMFile = "c:/ccc.ged"
' GEDCOMFile is the GEDCOM file in which all the people are in a tree.

'************************************
'* Globally define all the needed variables,
'* constants and arrays.
'*
'* The critical element for fast lookups is being able to
'* use an index-value lookup in an array.
'************************************
Const RINLimit = 100000, MRINLimit = 50000, FPLimit = 300, PFLimit = 200
' See Table1KidLimit in subroutine for limiting number of children and also the related array size.

Dim RinFlag(1 To 2, 1 To RINLimit) As Boolean ' Cell Value is TRUE or FALSE
Dim RinName(1 To RINLimit) As String ' Cell Value is the Person's name
Dim RinatDNA(1 To RINLimit) As Boolean ' Cell Value is the Person's atDNA-tested status
Dim NRINS As Integer ' Count of the number of RINs in the file
Dim MAXRIN As Integer ' The highest RIN number in the file
Dim Rin As Integer
Dim MRin As Integer



'************************************
'FamPers is a table of all the persons in a specific family.
'Each person (RIN) is in a separate record for the family (MRIN).
'So, a family is a collection of (MRIN, RIN) pairs in the array.
'************************************
' FamPers (or just FP) uses a family MRIN to find all members' RINs
Dim FamPers(1 To MRINLimit, 1 To FPLimit) As Long ' Cell Value is a +/- RIN
'************************************
'PersFam is a table of all the families to which a specific person belongs.
'Each family (MRIN) is connected to each child (RIN) in the family.
'So, a person is a collection of (RIN, MRIN) pairs in the array.
'************************************
```

```
' PersFam (of just PF) uses a person's RIN to find all MRINs to which he/she belongs
Dim PersFam(1 To RINLimit, 1 To PFLimit) As Long ' Cell Value is +/- MRIN
' + = Spouse (i.e., if +MRin or +Rin, then person is a spouse in that MRin)
' - = Child (i.e., if -MRin or -Rin, then person is a child in that MRin)

Dim NMRINS As Integer ' Count of the number of MRINs in the file
Dim MAXMRIN As Integer ' The highest MRIN number in the file

Dim FPRin As Long
Dim PFMRin As Long
Dim TargetLDNACoverage As Double
Dim TargetUDNACoverage As Double

Dim FPSpouse As Boolean
Dim PFSpouse As Boolean

Dim LoopCt As Long
Dim LoopCt2 As Long
Dim XYCt As Integer

Dim RightSide As String
Dim LeftSide As String

'The following are for writing the results out.
 Dim LineCount As Integer ' Line Counter for output report
 Dim OutputRow(1 To 500000) As String ' This can be an implicit limit on the number of children of any
given parent. <<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
 Dim OutputValue As String
'*************************
'* END OF DEFINITIONS
'*************************

Sub DNACoverage()
'*************************
'* START OF MAIN PROGRAM *
'*************************

'*****************************************************************************************
****************************
'*                        Click in this area, and
'*                        select RUN from the menu.
'*****************************************************************************************
****************************

StartTime = Time
Call InitVars ' Initialize the variables, so that restarts run correctly
Call ReadGEDCOM ' Read the GEDCOM File and assign all the variables.
PrepEndTime = Time
```

```
showit = "DONE WITH PREP--Start-" & StartTime & " PrepEnd-" & PrepEndTime & " NRINS =" & NRINS &
" MAXRIN =" & MAXRIN & " NMRINS =" & NMRINS & " MAXMRIN =" & MAXMRIN
MsgBox showit

'*************************************************************************************
****************************
' Write the first output line to the OutputRow array and increment the line counter.
'*************************************************************************************
****************************

LineCount = 1
OutputRow(LineCount) = "Calculation of the DNA Coverage of RIN " & XPerson & " (" &
RinName(XPerson) & ")"
LineCount = LineCount + 1

'*************************************************************************************
****************************
' Calculate the DNA Coverage of the target person.
'*************************************************************************************
****************************

TargetLDNACoverage = -1
TargetLDNACoverage = CalcLDNACoverage(XPerson) ' This is the real engine of the program for lower
bound calculation.
showit = "DONE WITH Lower Bound"
MsgBox showit


TargetUDNACoverage = -1
TargetUDNACoverage = CalcUDNACoverage(XPerson) ' This is the real engine of the program for lower
bound calculation.
showit = "DONE WITH Upper Bound"
MsgBox showit
showit = "1-LineCount=" & LineCount
MsgBox showit

'*************************************************************************************
****************************
' Now that the DNA Coverage is calculated, write all the lines of output to the Results Report worksheet.
'*************************************************************************************
****************************

Set wout = Sheets("Results Report")
wout.Activate
wout.Cells.ClearContents

' Writing output now
showit = "2-LineCount=" & LineCount
```

```
MsgBox showit
For j = 1 To LineCount
wout.Cells(j, 1) = OutputRow(j)
Next ' For j = 1 to LineCount

RunEndTime = Time
showit = "DONE"
MsgBox showit


'*************************
'* END OF MAIN PROGRAM *
'*************************
End Sub
Function InitVars()

NRINS = 0
MAXRIN = 0
NMRINS = 0
MAXMRIN = 0

For LoopCt = 1 To RINLimit
RinName(LoopCt) = ""
RinatDNA(LoopCt) = False ' Intialize the atDNA-test status as False for everyone

For LoopCt2 = 1 To PFLimit
PersFam(LoopCt, LoopCt2) = 0
Next ' For LoopCt2 = 1 To PFLimit
If LoopCt < MRINLimit + 1 Then
For LoopCt2 = 1 To FPLimit
FamPers(LoopCt, LoopCt2) = 0
Next ' For LoopCt2 = 1 To FPLimit
End If ' If LoopCt < MRINLimit + 1 Then
Next ' For LoopCt = 1 To RINLimit

UnableToExpandLevel = False
FoundAConnection = False

End Function
Function ReadGEDCOM()

'*************************************
'* Read through the GEDCOM file and
'* 1. Define RIN-Name Index -- OBSOLETE: and X and Y RIN Flag Arrays
'* 2. Define Family-Person File and Family Counter Array
'* Only the INDI records of the GEDCOM file are used.
'*************************************
```

```
Const ForReading = 1, ForWriting = 2, ForAppending = 3
Const TristateUseDefault = -2, TristateTrue = -1, TristateFalse = 0
Dim fso, ts, s

Set fso = CreateObject("Scripting.FileSystemObject")
Set ts = fso.OpenTextFile(GEDCOMFile, ForReading, TristateUseDefault)

Dim RINSDone As Boolean
RINSDone = False
Dim EndofINDI As Boolean
EndofINDI = False
Dim PersCount As Integer ' Will run from 1 to FPLimit
Dim FamCount As Integer ' Will run from 1 to PFLimit

'Dim loopct As Integer
'For loopct = 1 To 500

' *
' ***************** Start Main GEDCOM Read Routine **********************
' *

Do While ts.AtEndOfLine <> True ' Read all the records in the GEDCOM file
s = ts.ReadLine ' This reads the current line of the GEDCOM file and assigns it to the variable s.

' *
' ***************** Start Section FoundZeroINDI **********************
' * Find a "0 ... INDI" record which is the first record of Person X.
' * Capture the RIN number, the name, the atDNA flag (if any) and the records of the families to which
Person X belongs.
' * Remember that there are "0" records before the first "0 ... INDI". You have to bypass those first.
' *
FoundZeroINDI:
If RINSDone = False Then ' Keep processing until reach the end of the RINs. After that do the MRINs.
If Left(s, 1) = 0 Then ' Look for a zero in column 1.
If Right(s, 4) = "INDI" Then ' Found the initial record for a new person **********************
' This begins a lot of processing before you find the associated End If
NRINS = NRINS + 1 ' Increment the count of the number of RINs in the file

' FoundZeroINDI-Part 1 ***** Now find and capture the RIN value itself **********************
RightSide = Right(s, Len(s) - 4)
For ctspaces = 1 To 10 ' Find the first @ symbol that defines the RIN number location.
If Mid(RightSide, ctspaces, 1) = "@" Then
GoTo FoundRinAt
End If
Next ' For ctspaces = 1 To 10

FoundRinAt:
' Once a RIN is found, capture the RIN value
```

```
Rin = (Left(RightSide, (ctspaces - 1)))

' FoundZeroINDI-Part 2 ***** Now read the next line which is the name line and capture the name
*********************
s = ts.ReadLine
RinName(Rin) = Right(s, Len(s) - 7) ' Capture the name for Person X (the Person with this RIN).

' *
' FoundZeroINDI-Part 3 ******* Now, READ ALL THE REST OF THE RECORDS FOR Person X. ********
' *
' * Find any TYPE or FAMS or FAMC records for this person and load to arrays. ********
' "2 TYPE atDNA" records are those that indicate that the person has done an autosomal DNA test.
' "1 FAMS" records are those where the person is a spouse/parent in the family.
' "1 FAMC" records are those where the person is a child in the family.
' Do not confuse the "1 FAM" records of the person (RIN) with the "0 @F1@ FAM" MRIN records.
' The TYPE record, if there is one, comes before the FAMS or FAMC records.
' So, the initial search is for a TYPE record.
' If you find a new zero-record, then you have found a new person so that you have to start work on that
person.

EndofINDI = False
Do Until EndofINDI = True ' Read all the records of person X until done with person X.
s = ts.ReadLine
' Sending control to LoopToFindFAM causes this Do Loop to end processing for Person X and read the
next GEDCOM record.

' First, check to see if you have finished processing Person X.
If Left(s, 1) = 0 Then ' Check to see if you found a new 0 (zero) record which means you are done
processing Person X. So,stop processing person X.
EndofINDI = True ' You stop processing Person X by setting EndofINDI to TRUE to end the Do loop which
will take you to EndofINDI.
End If ' If Left(s, 1) = 0


'
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
<<<<<<<<<<<<<<<<<<<<<<<<<<<<
' *
' FoundZeroINDI-Part 3A ******* See if Person X did an autosomal DNA test. ********
' *
' You will know that they tested if they have a "2 TYPE atDNA" record.
'
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
<<<<<<<<<<<<<<<<<<<<<<<<<<<<
If Left(s, 12) = "2 TYPE atDNA" Then ' The person has tested for autosomal DNA. So, set their RinatDNA
flag to TRUE.
RinatDNA(Rin) = True
GoTo LoopToFindFAM ' Since you have processed this record, go on to read the next record.
```

```
End If ' If Left(s, 12) = "2 TYPE atDNA"
' *
' If you have reached this point, then you know that there is no "2 TYPE atDNA" left to find for Person X.
' Either there was one and you already processed it or else there was none.
' So, you now are looking for "1 FAMS" or "1 FAMC" records for Person X.
' *



'
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
<<<<<<<<<<<<<<<<<<<<<<<<<<<<
' *
'      FoundZeroINDI-Part 3B ******* Find FAMS and/or FAMC records of Person X. ********
' *
'
<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
<<<<<<<<<<<<<<<<<<<<<<<<<<<<
If Left(s, 5) <> "1 FAM" Then ' If you do not have a FAMS or FAMC record, go on to the next GEDCOM
record.
GoTo LoopToFindFAM ' Since this record is not relevant, go on to read the next GEDCOM record.
End If ' If Left(s, 5) <> "1 FAM"
' If you reach this point, you have found a "1 FAM" record. So check to see if it is FAMS or FAMC.

' *******************************************************************************
'      FoundZeroINDI-Part 3B1: FAMS Check and Processing
' *******************************************************************************
If Mid(s, 3, 4) = "FAMS" Then ' Found the record of the family in which Person X is a parent/spouse.
' Write as POSITIVE number since positive numbers will designate parents and negatives will designate
children in the family.
' Parse to find the S-MRIN number.
'
' The following code starts at the 9th position in the record and searches for @.
' The record is "1 FAMS $F" in the first 9 positions. The MRIN value starts in the 10th position.
' To find the MRIN value, you have to find the second @ in the record.
' You then know the starting and ending positions of the MRIN value so that you can capture it.
RightSide = Right(s, Len(s) - 9)
For ctspaces = 1 To 10
If Mid(RightSide, ctspaces, 1) = "@" Then
GoTo FoundSMRinAt ' FoundSMRinAt is the subroutine that captures the value of the MRIN in which
Person X is a spouse/parent.
End If ' If Mid(RightSide, ctspaces, 1) = "@"
Next ' For ctspaces = 1 To 10
' *
' Subroutine FoundSMRinAt to capture the FAMS MRIN
' * There is a hazard here. If any GEDCOM ever has a huge MRIN number, then the processing will fall
through here.
' * That would lead to the subroutine FoundSMRinAt being executed when it should not.
```

```
' *
FoundSMRinAt:
MRin = (Left(RightSide, (ctspaces - 1))) ' The MRIN to be captured starts in the right side 1st position (column 10) and ends in the position just before ctspaces.
' Store in the Family-Person Array
For PersCount = 1 To FPLimit
If FamPers(MRin, PersCount) = 0 Then
FamPers(MRin, PersCount) = Rin
GoTo LoadFAMStoPF
End If
' Since this Rin slot is already filled, check the next one.
Next ' For PersCount = 1 To 30
LoadFAMStoPF:
' Store in the Person-Family Array
For FamCount = 1 To PFLimit
If PersFam(Rin, FamCount) = 0 Then
PersFam(Rin, FamCount) = MRin
' showit = "PersFam(" & Rin & "," & FamCount & ") = " & MRin
' MsgBox showit
GoTo LoopToFindFAM ' done processing this FAMS record
End If
' Since this Rin slot is already filled, check the next one.
Next ' For FamCount = 1 To 30
End If ' If Right(s, 4) = "FAMS"

' ************************************************************************
'       FoundZeroINDI-Part 3B2: FAMC Check and Processing
' ************************************************************************
If Mid(s, 3, 4) = "FAMC" Then
' Write as NEGATIVE number since positive numbers will designate parents and negatives will designate children in the family.
' Parse the C-MRIN.
'
' The following code starts at the 9th position in the record and searches for @.
' The record is "1 FAMC $F" in the first 9 positions. The MRIN value starts in the 10th position.
' To find the MRIN value, you have to find the second @ in the record.
' You then know the starting and ending positions of the MRIN value so that you can capture it.
RightSide = Right(s, Len(s) - 9)
For ctspaces = 1 To 10
If Mid(RightSide, ctspaces, 1) = "@" Then
GoTo FoundCMRinAt
End If
Next
' *
' Subroutine FoundCMRinAt to capture the FAMC MRIN
' * There is a hazard here. If any GEDCOM ever has a huge MRIN number, then the processing will fall through here.
' * That would lead to the subroutine FoundCMRinAt being executed when it should not.
```

```
' *
FoundCMRinAt:
MRin = (Left(RightSide, (ctspaces - 1))) ' The MRIN to be captured starts in the right side 1st position
(column 10) and ends in the position just before ctspaces.
' Store in the Family-Person Array
For PersCount = 1 To FPLimit
If FamPers(MRin, PersCount) = 0 Then
FamPers(MRin, PersCount) = -Rin
GoTo LoadFAMCtoPF
End If
' Since this Rin slot is already filled, check the next one.
Next ' For PersCount = 1 To 30
LoadFAMCtoPF:
' Store in the Person-Family Array
For FamCount = 1 To PFLimit
If PersFam(Rin, FamCount) = 0 Then
PersFam(Rin, FamCount) = -MRin
GoTo LoopToFindFAM ' done processing this FAMC record
End If
' Since this Rin slot is already filled, check the next one.
Next ' For FamCount = 1 To 30
End If ' If Right(s, 4) = "FAMC"



LoopToFindFAM:
' This step takes you to the next record in the GEDCOM.
Loop ' Do Until EndofINDI = True
'
*****************************************************************************************
************
'       End of FoundZeroINDI-Part 3 and of all processing for Person X.
'
*****************************************************************************************
************




'
*****************************************************************************************
************
'   EndofINDI: Finished processing Person X. So, see if it current record is the last RIN record.
'
*****************************************************************************************
************
EndofINDI: ' You have finished the current individual's records.
' The first MRIN record will have the form "0 ... FAM".
```

```
' A search for "FAM" in the rightmost 3 characeters will tell you that you have reached the MRIN
reocrds.
If Right(s, 3) = "FAM" Then ' The first "0 ... FAM" record in the GEDCOM indicates RINS are done.
RINSDone = True
GoTo RINSDone
End If ' If Right(s, 3) = "FAM"
GoTo FoundZeroINDI: ' go back and process the new individual you have found.
Else ' If Right(s, 4) = "INDI" Then
If Right(s, 3) = "FAM" Then ' The first "0 ... FAM" indicates RINS are done.
RINSDone = True
GoTo RINSDone
End If ' If Right(s, 3) = "FAM" Then
End If ' If Right(s, 4) = "INDI" Then
End If ' If Left(s, 1) = 0 Then
Else ' If RINSDone = False Then

End If ' If RINSDone = False Then


'
*************************************************************************************
************
'                                RINSDone
'              You have finished processing all the RINs in the GEDCOM.
'              Capture the NMAXRINS count of RINs and MAXMRIN count of MRINs.
'                        Then exit the ReadGEDCOM Function.
'
*************************************************************************************
************
RINSDone:

'Next ' For loopct = 1 To 100
Loop 'Do While ts.AtEndOfLine <> True

ts.Close

MAXRIN = Rin ' Capture the highest RIN, which will be the last one
' Figure out MAXMRIN and NMRINS
For LoopCt = 1 To MRINLimit
If FamPers(LoopCt, 1) <> 0 Then
NMRINS = NMRINS + 1
MAXMRIN = LoopCt
End If ' If FamPers(LoopCt, 1) <> 0
Next ' For LoopCt = 1 To MRINLimit

End Function ' ReadGEDCOM()
Function CalcLDNACoverage(ReceivedRIN)
```

```
'Initial setup for Target Person
'Find all children of target person and create upper and lower bounds arrays for them --
LTable2(ChildSeq,ChildRIN,ChildCoverage) and UTable2
'Set values of N, P, M and set up LTable1 and LTable2 of Venn Diagram representation for lower and
upper bounds
'For each child, call GetChildCoverage for child and enter it in Table2
'When processed all children, use tables to calculate target person's coverage, report it and end.

'CalcLDNACoverage:
'Find all children of received person and create upper and lower bounds arrays for them --
LTable2(ChildSeq,ChildRIN,ChildCoverage) and UTable2
'Set values of N, P, M and set up LTable1 and LTable2 of Venn Diagram representation for lower and
upper bounds
'For each child, call GetChildCoverage for child and enter it in Table2
'When processed all children, use tables to calculate received person's coverage, report it for family,
return the coverage to call and end.

CalcLDNACoverage = 0
Dim RinReceived As Integer
RinReceived = ReceivedRIN

Dim MRINFound As Integer
Dim KidsFound As Integer

'*************************************************************
'  Set up LTable2 and UTable2 (Lower and Upper-Bound Table 2s)
'  The index value is the number of the child in the table.
'  xTable2RIN is the RIN of the child.
'  xTable2Coverage is the DNA coverage of the child, initially zero.
'*************************************************************
Dim LTable2RIN(1 To FPLimit) As Integer
Dim LTable2Coverage(1 To FPLimit) As Double

'*************************************************************
'  If the person has DNA-tested, set value to 1 (100%).
'  Then exit the function.
'*************************************************************
If RinatDNA(RinReceived) = True Then ' Do I need to write an outputrow for
this???????????????????????????????????????????????????????
   CalcLDNACoverage = 1
   GoTo EndCalc ' Go to the exit of this function.
End If

'*************************************************************************************
************
'       If you reach this point, the person has NOT done an autosomal DNA test.
'          So, find Person X's children to calculate the person's coverage.
```

```
'*************************************************************************************
************

'*************************************************************************************
************
' To find the children, first find all the PersFam entries for which person X is a parent.
' That is, find all the families in which person X is a parent
' Search PersFam for the RIN of Person X (the row variable of PersFam).
' Then for each child of that MRIN (in FamPers), if the value (the RIN) is positive, Person X is a parent in
that MRIN.
' So, use that MRIN to search FamPers.
' For any FamPers record for this MRIN, if the second element (RIN) is negative, then that person is a
child of that parent in that family.
' So, capture the postive value of that child's RIN into Table 2 (both) as a child of Person X.
'*************************************************************************************
************
KidsFound = 0
For CountFamilies = 1 To PFLimit
   If PersFam(RinReceived, CountFamilies) > 0 Then ' Have found a family record (MRIN) of Person X - so
now find all children of that family and load to Table 2
      MRINFound = PersFam(RinReceived, CountFamilies)
      For CountChildren = 1 To FPLimit
         If FamPers(MRINFound, CountChildren) < 0 Then
            KidsFound = KidsFound + 1
            LTable2RIN(KidsFound) = Abs(FamPers(MRINFound, CountChildren))
            LTable2Coverage(KidsFound) = 0 ' Initialize the Child's DNA Coverage at zero
         End If ' If FamPers(MRINFound, CountChildren) <> 0
      Next ' For CountChildren = 1 To FPLimit
   End If ' If PersFam(RinReceived, CountFamilies) <> 0
Next ' For CountFamilies = 1 To PFLimit

'*************************************************************************************
************
' So now, you have built both versions (L and U) of Table2.
' So calculate the key constants.
' N = number of children of the parent
' P = Pieces of the Venn diagram of the parent =(2^N)-1
' M = Max percent of parent that each piece of the Venn diagram can contribute = 1/(2^N)
' M is calculated as a decimal and not as a percent
'*************************************************************************************
************
Dim N, P As Long
Dim M As Double

N = KidsFound
P = (2 ^ N) - 1
M = 1 / (2 ^ N)
```

```
'*************************************************************************************
************
' Now create both versions (L and U) of Table1. Each column is a separate 1-dim array.
' Each row of Table 1 has one column for each child with the binary value of that row number in those
columns.
' Each row of Table 1 also has one column for M which is the constant M for all rows.
' Each row of Table 1 also has one column for W (weight)/
' -- For the lower bound W is the maximum coverage of any child with a 1 in their column in that row.
'*************************************************************************************
************
' -- For the upper bound W is the minimum of 1 or of the sum of the coverages of all the children with a
1 in their column in that row.
'*************************************************************************************
************
' Each row of Table 1 also has one column for R (result) which is M * W.
'*************************************************************************************
************

'*************************************************************
'   Set up LTable1 and UTable1 (Lower and Upper-Bound Table 1s)
'   The index value is the number of the piece of the Venn diagram in the table.
'   Thus the index number is the row number in Table1.
'
'*************************************************************
Const Table1PieceLimit = 4096 ' Allows for up to 12 children
Const Table1KidLimit = 12 ' Allows for up to 12 children

Dim LTable1Binary(1 To Table1PieceLimit, 1 To Table1KidLimit) As Integer
Dim LTable1M(1 To Table1PieceLimit) As Double
Dim LTable1W(1 To Table1PieceLimit) As Double
Dim LTable1R(1 To Table1PieceLimit) As Double

'*************************************************************
'   Set the Binary Value for Each Piece/Row in the Child Columns
'   Convert Venn Diagragm piece number to its binary equivalent and put bits into Table1 child cells
'   Works by reducing piece number by power of 2 on each iteration
'*************************************************************
For PieceNum = 1 To P
   LTable1W(PieceNum) = 0
   LTable1M(PieceNum) = M
   LTable1R(PieceNum) = 0
   Remainder = PieceNum
   For KidNum = 1 To N
      BitPower = N - KidNum
      Bit = WorksheetFunction.Power(2, BitPower)  ' Power(2, BitPower)
      If Remainder >= Bit Then
         LTable1Binary(PieceNum, KidNum) = 1
         Remainder = Remainder - Bit
```

```
      Else
         LTable1Binary(PieceNum, KidNum) = 0
      End If ' If Remainder >= Bit
   Next ' For KidNum = 1 To N
Next ' For PieceNum = 1 To P

'*************************************************************************************
************
' Now that both tables are set up, step through each child to obtain their DNA coverage in Table2.
' This is done be recursive calls to this same routine.
'*************************************************************************************
************
For ChildCount = 1 To N
   ChildRIN = LTable2RIN(ChildCount)
   TargetLDNACoverage = CalcLDNACoverage(ChildRIN)
   LTable2Coverage(ChildCount) = TargetLDNACoverage
Next ' For ChildCount = 1 To N

'+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
'+++++++++++++++++++++++++++++++++++++++++++++++++ THE KEY CALCULATION
++++++++++++++++++++++++++++++++++++++++++++++++++
'+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
' Now the children have all been found and their own DNA Coverage has been calculated.
' So, calculate the coverage of their parent.
' This is the KEY CALCULATION of this function.
' It is the only place that the calculation of the lower and upper bounds differ.
'+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
'+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
'+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++

'*************************************************************************************
************
' Now that the DNA coverage of all children is known, calculate the DNA coverage of Person X who is
RINReceived.
' Calcualte the contribution of each piece of the Venn Diagram and add them up.
''*************************************************************************************
************
Dim MaxChildDNA As Double
MaxChildDNA = 0
Dim ChildProduct As Double
Dim WorkingCalcLDNACoverage As Double
WorkingCalcLDNACoverage = 0
```

```
For PieceCount = 1 To P
'
*********************************************************************************
********
' Calculate lower bound W = max of the products of each child's coverage (from Table 2)
' Calculate lower bound R = W * M
'
*********************************************************************************
********
   MaxChildDNA = 0
   For ChildCount = 1 To N
   '
*********************************************************************************
********
   ' Calculate lower bound W = max of the products of each child's coverage (from Table 2)
   '
*********************************************************************************
********
      ChildProduct = LTable2Coverage(ChildCount) * LTable1Binary(PieceCount, ChildCount)
      MaxChildDNA = WorksheetFunction.Max(MaxChildDNA, ChildProduct)
      LTable1W(PieceCount) = MaxChildDNA
   Next 'For ChildCount = 1 To N
   LTable1R(PieceCount) = LTable1W(PieceCount) * LTable1M(PieceCount)
   WorkingCalcLDNACoverage = WorkingCalcLDNACoverage + LTable1R(PieceCount)
Next ' For PieceCount = 1 To P

'
*********************************************************************************
********
' Set the output value to be returned
'
*********************************************************************************
********
CalcLDNACoverage = WorkingCalcLDNACoverage

'
*********************************************************************************
********
' Now that the calculation for this person is complete, write the output rows
' for this person and the children and Tables 1 and 2 values.
'
*********************************************************************************
********
OutputValue = "-------------------------------------------------------------------"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1
```

```
OutputValue = "LOWER BOUND DNA Coverage of RIN " & RinReceived & " (" & RinName(RinReceived) &
") = " & CalcLDNACoverage
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

OutputValue = "Children of RIN " & RinReceived & " (" & RinName(RinReceived) & ")"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

For ChildCount = 1 To N
   OutputValue = "RIN " & LTable2RIN(ChildCount) & " (" & RinName(LTable2RIN(ChildCount)) & ")"
   OutputRow(LineCount) = OutputValue
   LineCount = LineCount + 1
Next ' For ChildCount = 1 to N

'
********************************************************************************
********
' Write out Table 2
'
********************************************************************************
********
OutputValue = "***** Table 2 *****"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

OutputValueRIN = ""
OutputValueCoverage = ""
For ChildCount = 1 To N
   OutputValueRIN = OutputValueRIN & LTable2RIN(ChildCount) & "--"
   OutputValueCoverage = OutputValueCoverage & LTable2Coverage(ChildCount) & "--"
Next ' For ChildCount = 1 to N

OutputRow(LineCount) = "RINS--" & OutputValueRIN
LineCount = LineCount + 1

OutputRow(LineCount) = "DNA--" & OutputValueCoverage
LineCount = LineCount + 1

'
********************************************************************************
********
' Write out Table 1
'
********************************************************************************
********
OutputValue = "***** Table 1 *****"
OutputRow(LineCount) = OutputValue
```

```
LineCount = LineCount + 1

OutputValuePiece = ""
OutputValueHeader = "----R----M----W<<<<"
For ChildCount = 1 To N
   OutputValueHeader = OutputValueHeader & LTable2RIN(ChildCount) & "+"
Next ' For ChildCount = 1 To N

OutputRow(LineCount) = OutputValueHeader
LineCount = LineCount + 1

OutputValuePiece = ""
For PieceCount = 1 To P
   OutputValuePiece = OutputValuePiece & "----" & LTable1R(PieceCount) & "--" &
LTable1M(PieceCount) & "--" & LTable1W(PieceCount) & "<<<<"
   For ChildCount = 1 To N
      OutputValuePiece = OutputValuePiece & LTable1Binary(PieceCount, ChildCount) & "+"
   Next ' For ChildCount = 1 to N
   OutputRow(LineCount) = OutputValuePiece
   LineCount = LineCount + 1
   OutputValuePiece = ""
Next ' For PieceCount = 1 to P
EndCalc:
End Function ' CalcLDNACoverage
Function CalcUDNACoverage(ReceivedRIN)
'++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+++'

'Initial setup for Target Person
'Find all children of target person and create upper and lower bounds arrays for them --
LTable2(ChildSeq,ChildRIN,ChildCoverage) and UTable2
'Set values of N, P, M and set up LTable1 and LTable2 of Venn Diagram representation for lower and
upper bounds
'For each child, call GetChildCoverage for child and enter it in Table2
'When processed all children, use tables to calculate target person's coverage, report it and end.

'CalcLDNACoverage:
'Find all children of received person and create upper and lower bounds arrays for them --
LTable2(ChildSeq,ChildRIN,ChildCoverage) and UTable2
'Set values of N, P, M and set up LTable1 and LTable2 of Venn Diagram representation for lower and
upper bounds
'For each child, call GetChildCoverage for child and enter it in Table2
'When processed all children, use tables to calculate received person's coverage, report it for family,
return the coverage to call and end.

CalcUDNACoverage = 0
Dim RinReceived As Integer
```

```
RinReceived = ReceivedRIN

Dim MRINFound As Integer
Dim KidsFound As Integer

'*************************************************************
'  Set up LTable2 and UTable2 (Lower and Upper-Bound Table 2s)
'  The index value is the number of the child in the table.
'  xTable2RIN is the RIN of the child.
'  xTable2Coverage is the DNA coverage of the child, initially zero.
'*************************************************************
Dim UTable2RIN(1 To FPLimit) As Integer
Dim UTable2Coverage(1 To FPLimit) As Double

'*************************************************************
'  If the person has DNA-tested, set value to 1 (100%).
'  Then exit the function.
'*************************************************************
If RinatDNA(RinReceived) = True Then ' Do I need to write an outputrow for
this??????????????????????????????????????????????????????
   CalcUDNACoverage = 1
   GoTo EndCalc ' Go to the exit of this function.
End If

'*************************************************************************************
************
'       If you reach this point, the person has NOT done an autosomal DNA test.
'         So, find Person X's children to calculate the person's coverage.

'*************************************************************************************
************

'*************************************************************************************
************
' To find the children, first find all the PersFam entries for which person X is a parent.
' That is, find all the families in which person X is a parent
' Search PersFam for the RIN of Person X (the row variable of PersFam).
' Then for each child of that MRIN (in FamPers), if the value (the RIN) is positive, Person X is a parent in
that MRIN.
' So, use that MRIN to search FamPers.
' For any FamPers record for this MRIN, if the second element (RIN) is negative, then that person is a
child of that parent in that family.
' So, capture the postive value of that child's RIN into Table 2 (both) as a child of Person X.
'*************************************************************************************
************
KidsFound = 0
For CountFamilies = 1 To PFLimit
```

```
   If PersFam(RinReceived, CountFamilies) > 0 Then ' Have found a family record (MRIN) of Person X - so now find all children of that family and load to Table 2
      MRINFound = PersFam(RinReceived, CountFamilies)
      For CountChildren = 1 To FPLimit
         If FamPers(MRINFound, CountChildren) < 0 Then
            KidsFound = KidsFound + 1
            UTable2RIN(KidsFound) = Abs(FamPers(MRINFound, CountChildren))
            UTable2Coverage(KidsFound) = 0 ' Initialize the Child's DNA Coverage at zero
         End If ' If FamPers(MRINFound, CountChildren) <> 0
      Next ' For CountChildren = 1 To FPLimit
   End If ' If PersFam(RinReceived, CountFamilies) <> 0
Next ' For CountFamilies = 1 To PFLimit

'*************************************************************************************
************
' So now, you have built both versions (L and U) of Table2.
' So calculate the key constants.
' N = number of children of the parent
' P = Pieces of the Venn diagram of the parent =(2^N)-1
' M = Max percent of parent that each piece of the Venn diagram can contribute = 1/(2^N)
' M is calculated as a decimal and not as a percent
'*************************************************************************************
************
Dim N, P As Long
Dim M As Double

N = KidsFound
P = (2 ^ N) - 1
M = 1 / (2 ^ N)



'*************************************************************************************
************
' Now create both versions (L and U) of Table1. Each column is a separate 1-dim array.
' Each row of Table 1 has one column for each child with the binary value of that row number in those columns.
' Each row of Table 1 also has one column for M which is the constant M for all rows.
' Each row of Table 1 also has one column for W (weight)/
' -- For the lower bound W is the maximum coverage of any child with a 1 in their column in that row.
'*************************************************************************************
************
' -- For the upper bound W is the minimum of 1 or of the sum of the coverages of all the children with a 1 in their column in that row.
'*************************************************************************************
************
' Each row of Table 1 also has one column for R (result) which is M * W.
```

```
'*************************************************************************************
************

'*************************************************************
'   Set up LTable1 and UTable1 (Lower and Upper-Bound Table 1s)
'   The index value is the number of the piece of the Venn diagram in the table.
'   Thus the index number is the row number in Table1.
'
'*************************************************************
Const Table1PieceLimit = 4096 ' Allows for up to 12 children
Const Table1KidLimit = 12 ' Allows for up to 12 children

Dim UTable1Binary(1 To Table1PieceLimit, 1 To Table1KidLimit) As Integer
Dim UTable1M(1 To Table1PieceLimit) As Double
Dim UTable1W(1 To Table1PieceLimit) As Double
Dim UTable1R(1 To Table1PieceLimit) As Double

'*************************************************************
'   Set the Binary Value for Each Piece/Row in the Child Columns
'   Convert Venn Diagragm piece number to its binary equivalent and put bits into Table1 child cells
'   Works by reducing piece number by power of 2 on each iteration
'*************************************************************

For PieceNum = 1 To P
   UTable1W(PieceNum) = 0
   UTable1M(PieceNum) = M
   UTable1R(PieceNum) = 0
   Remainder = PieceNum
   For KidNum = 1 To N
      BitPower = N - KidNum
      Bit = WorksheetFunction.Power(2, BitPower)  ' Power(2, BitPower)
      If Remainder >= Bit Then
         UTable1Binary(PieceNum, KidNum) = 1
         Remainder = Remainder - Bit
      Else
         UTable1Binary(PieceNum, KidNum) = 0
      End If ' If Remainder >= Bit
   Next ' For KidNum = 1 To N
Next ' For PieceNum = 1 To P

'*************************************************************************************
************
' Now that both tables are set up, step through each child to obtain their DNA coverage in Table2.
' This is done be recursive calls to this same routine.
'*************************************************************************************
************
For ChildCount = 1 To N
   ChildRIN = UTable2RIN(ChildCount)
```

```
   TargetUDNACoverage = CalcUDNACoverage(ChildRIN)
   UTable2Coverage(ChildCount) = TargetUDNACoverage
Next ' For ChildCount = 1 To N



'++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
'+++++++++++++++++++++++++++++++++++++++++++++++++ THE KEY CALCULATION
++++++++++++++++++++++++++++++++++++++++++++++++++
'++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
' Now the children have all been found and their own DNA Coverage has been calculated.
' So, calculate the coverage of their parent.
' This is the KEY CALCULATION of this function.
' It is the only place that the calculation of the lower and upper bounds differ.
'++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
'++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
'++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++
WorkingCalcUDNACoverage = 0
For PieceCount = 1 To P
'
*********************************************************************************************
********
' Calculate upper bound W = max of the products of each child's coverage (from Table 2)
' Calculate upper bound R = W * M
'
*********************************************************************************************
********
   SumChildDNA = 0
   For ChildCount = 1 To N
   '
*********************************************************************************************
********
   ' Calculate lower bound W = max of the products of each child's coverage (from Table 2)
   '
*********************************************************************************************
********
      ChildProduct = UTable2Coverage(ChildCount) * UTable1Binary(PieceCount, ChildCount) * M
      SumChildDNA = SumChildDNA + ChildProduct
      UTable1W(PieceCount) = SumChildDNA
   Next 'For ChildCount = 1 To N
   UTable1R(PieceCount) = WorksheetFunction.Min(UTable1W(PieceCount), UTable1M(PieceCount))
   WorkingCalcUDNACoverage = WorkingCalcUDNACoverage + UTable1R(PieceCount)
Next ' For PieceCount = 1 To P
```

```
'
*********************************************************************************
********
' Set the output value to be returned
'
*********************************************************************************
********
CalcUDNACoverage = WorkingCalcUDNACoverage

'
*********************************************************************************
********
' Now that the calculation for this person is complete, write the output rows
' for this person and the children and Tables 1 and 2 values.
'
*********************************************************************************
********
OutputValue =
"+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

OutputValue = "UPPER BOUND DNA Coverage of RIN " & RinReceived & " (" & RinName(RinReceived) &
") = " & CalcUDNACoverage
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

OutputValue = "Children of RIN " & RinReceived & " (" & RinName(RinReceived) & ")"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

For ChildCount = 1 To N
   OutputValue = "RIN " & UTable2RIN(ChildCount) & " (" & RinName(UTable2RIN(ChildCount)) & ")"
   OutputRow(LineCount) = OutputValue
   LineCount = LineCount + 1
Next ' For ChildCount = 1 to N


'
*********************************************************************************
********
' Write out Table 2
'
*********************************************************************************
********
```

```
OutputValue = "***** Table 2 *****"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

OutputValueRIN = ""
OutputValueCoverage = ""
For ChildCount = 1 To N
   OutputValueRIN = OutputValueRIN & UTable2RIN(ChildCount) & "--"
   OutputValueCoverage = OutputValueCoverage & UTable2Coverage(ChildCount) & "--"
Next ' For ChildCount = 1 to N

OutputRow(LineCount) = "RINS--" & OutputValueRIN
LineCount = LineCount + 1

OutputRow(LineCount) = "DNA--" & OutputValueCoverage
LineCount = LineCount + 1

'
********************************************************************************
********
' Write out Table 1
'
********************************************************************************
********
OutputValue = "***** Table 1 *****"
OutputRow(LineCount) = OutputValue
LineCount = LineCount + 1

OutputValuePiece = ""
OutputValueHeader = "----R----M----W<<<<"
For ChildCount = 1 To N
   OutputValueHeader = OutputValueHeader & UTable2RIN(ChildCount) & "+"
Next ' For ChildCount = 1 To N

OutputRow(LineCount) = OutputValueHeader
LineCount = LineCount + 1

OutputValuePiece = ""
For PieceCount = 1 To P
   OutputValuePiece = OutputValuePiece & "----" & UTable1R(PieceCount) & "--" &
UTable1M(PieceCount) & "--" & UTable1W(PieceCount) & "<<<<"
   For ChildCount = 1 To N
      OutputValuePiece = OutputValuePiece & UTable1Binary(PieceCount, ChildCount) & "+"
   Next ' For ChildCount = 1 to N
   OutputRow(LineCount) = OutputValuePiece
   LineCount = LineCount + 1
   OutputValuePiece = ""
Next ' For PieceCount = 1 to P
```

```
EndCalc:
End Function ' CalcLDNACoverage
```

Appendix C: Understanding the Visual Basic Program

While the program looks formidable, it really is quite simple. Unfortunately, the Visual Basic editor only allows for the addition of comment lines and also inserts underscore lines between sections but otherwise is primitive for providing understanding.

The program has these sections.

1. Declarations Section (the first section with no title)
2. Sub DNACoverage()
3. Function InitVars()
4. Function ReadGEDCOM()
5. Function CalcLDNACoverage(ReceivedRIN)
6. Function CalcUDNACoverage(ReceivedRIN)

Do not consider this an exemplar of ideal programming. It is a quick (as quick as the primitive Visual Basic editor and error messages allow) and a bit dirty program that could probably be cleaned up a lot. It is not at all state-of-the-art.

1-Declarations Section

The declarations section sets up all the "global" constants, variables and arrays. "Global" means that these objects can be used in any part of the program. A sub-function (such as the one that reads the GEDCOM file) can use "local" constants, variables and arrays that cannot be used in other sections of the program.

2-Sub DNACoverage()

Think of this section as the orchestra conductor. A simple line here can actually be an instruction to a sub-function than can lead to a great deal of computation. While there is a fair amount of housekeeping going on in the section for the eventual reporting of the results, it really does five basic things.

1. Initiate the variables that will be used to store the program's internal representation of the GEDCOM file.
2. Read the GEDCOM file and store it into the internal representation.
3. Calculate the lower bound coverage of the target person.
4. Calculate the upper bound coverage of the target person.
5. Print the results report on the "Results Report" worksheet.

3-Function InitVars()

This brief section simply initializes the internal representation of the GEDCOM file as not knowing anything. This ensures that there are no faulty conclusions reached at any later decision point.

4-Function ReadGEDCOM()

This routine reads the GEDCOM and internally stores the information on the people and their relationships and which ones have DNA-tested.

5-Function CalcLDNACoverage(ReceivedRIN)

This is one of the two sections that are the real engines of the program: this one that calculates the lower bound and the other one that calculates the upper bound. These are the recursive sections which call more executions of themselves as the processing burrows down through the family tree and then back up to calculate the DNA coverage of a single parent from the coverage of all their children who have DNA coverage.

The two functions (lower and upper bound) are identical except for the step in which all the children's coverages are combined to calculate the parent's coverage.

The function receives the RIN of the person whose DNA coverage is to be calculated. It then finds all the children of that parent in the GEDCOM file (as internally represented). The function sets up the internal Table 1 representation of the Venn diagram for this parent and their children.

The function then steps through the children one by one. If a child has tested, their coverage is set at 100%, but if not then the function calls itself with the RIN of the child as the person whose DNA coverage is to be calculated.

6-Function CalcUDNACoverage(ReceivedRIN)

This is the other shoe in the pair. This function calculates the upper bound of the DNA coverage of the designated person from the coverage of his/her children. It is identical to the lower bound step except in how it calculates the parent's coverage once the coverage of all the children is known. See the main part of this paper for the specifics of how those two different coverages are calculated.

---

[i] Journal of Genetic Genealogy, vol. 10, no. 1, Fall 2022.

[ii] My real purpose is that all family tree software will someday have autosomal DNA coverage as a standard feature, just as we have relationship calculation and other tools.

[iii] The Hurricane Path "Cone of Uncertainty" Analogy: The "cone of uncertainty" in hurricane path predictions provides a useful analogy. Different models forecast different paths that when all shown on the same map look like different strands of spaghetti noodles. The cone of uncertainty holds all those different predicted paths. In the same way, the range of DNA coverage holds all possible combinations that the descendant DNA tests can provide for the coverage of the target person. We cannot know in a theoretical model just which is the actual combination possible from those DNA tests. Once the raw DNA results for all the tests are combined, they will define a single combination. But with only the theoretical and not the actual results, we can only calculate a range of possible combinations.

[iv] For cases where there are no generations for which there are no tested children, there is no range so that a single number can represent the DNA coverage. In such cases, the average and both the lower and upper bounds are all the same. This situation only exists in theoretical models. When actual DNA segments are known and can be included in the calculation, there is no longer uncertainty. Thus, the DNA reconstruction in GFG or in Borland Genetics will result in a specific percent of coverage based on the actual DNA results and not on a theoretical model that only provides estimates.

Simulations can also give a sense of reality. In hurricane forecasting, these are the "spaghetti" charts that align the hurricane paths predicted by different models. The different paths give a sense of the size of the cone of

uncertainty. In the same way, running many simulations of actual family trees give a strong sense of the size of the range of DNA coverage. The simulations of Amy Williams and the broader complex family tree simulations of Briton Nicholson explored this. Even for recent connections where the theoretical models can provide a single estimate, the simulations result in a range of possibilities with the theoretical model's calculated coverage usually right in the middle. The benefit of the theoretical model for ranges is that it can definitively calculate the upper and lower bounds so that the full possibility of the range of coverages can be known with certainty (within the limits covered in Appendix A).

[v] In a lineage-linked relational database, these two operations (top-down tracing of the descendants and bottom-up propagation of the DNA coverage) work in parallel in the same call of the sub-routine. In the Graphs for Genealogists implementation, Dr. David Stumpf leveraged the power of the graph database to replace the top-down tracing of the descendants with database queries that created directed graphs connecting the target ancestor with each tested descendant.

[vi] The only estimates that a recursive algorithm can accurately calculate are the lower and upper bound. Attempting to propagate any other number (such as the average) within the range inevitably leads to distortion, either toward over or under estimation. The distortion is minimal for fewer children but becomes significantly greater with more children if the children do not have high DNA coverage themselves.

[vii] https://dnapainter.com/tools/sharedcmv4

[viii] Nicholson, Briton. "The Overused CentiMorgan". Dec. 31, 2020. (https://dna-sci.com/2020/12/31/the-overused-centimorgan)

[ix] Nicholson, Briton. "Modeling the Inheritance of Ancestral DNA". Feb. 8, 2020. (https://dna-sci.com/2020/02/08/modeling-the-inheritance-of-ancestral-dna)

[x] Borland, Kevin at https://www.facebook.com/groups/borlandgenetics/posts/404094730131696

[xi] Stumpf, Dr. David at https://www.wai.md/post/ancestor-reconstruction