JoGG Journal of Genetic Genealogy

Volume 8, Number 1 ISSN 1557-3796 Fall, 2016

Columns

Reports

Editorials

Reviews and Announcements

Editor's Corner

The editorial board of the *Journal of Genetic Genealogy* is pleased to present you with our latest issue, which includes a case study by Joseph Fox III and David Fox that models the use of yDNA in conjunction with a paper trail to assess family trees; a scientific research paper by Doron Yacobi and Felice Bedford with mtDNA evidence for intermarriage between Ashkenazi Jews and non-Jewish Europeans early in the Jewish settlement of Europe; editorials by CeCe Moore and Blaine Bettinger; book and product reviews; and our regular 'Satiable Curiosity column by Ann Turner.

The genetic genealogy community has benefitted enormously from significant advances in recent years. At the time that our last issue was published in 2011, yDNA STR data and mtDNA sequencing were the prevalent types of genealogical data available. Autosomal DNA for genealogical purposes was in its infancy.

Since that time, atDNA testing has exploded in popularity among genealogists and laypeople — who might be mainly interested in their ethnicity profiles or health information — alike. This growth is evident across all three of the major genealogical testing companies as well as National Geographic's Genographic project (Figure 1). Also new to the genetic genealogy community since our last issue are targeted SNP testing from YSEQ and the Big YTM test (Family Tree DNA).

These technological advances expand the opportunities for academic and citizen scientists alike to contribute to the field of genetic genealogy. To that end, we are proud to revive JoGG as a venue for communication and the exchange of new ideas. Prospective authors should consult our Instructions for Authors and can submit their completed manuscripts to jogg@isogg.org. We are also actively recruiting volunteer peer reviewers, copy editors, and layout people.

Leah Larkin, Ph.D.
Editor

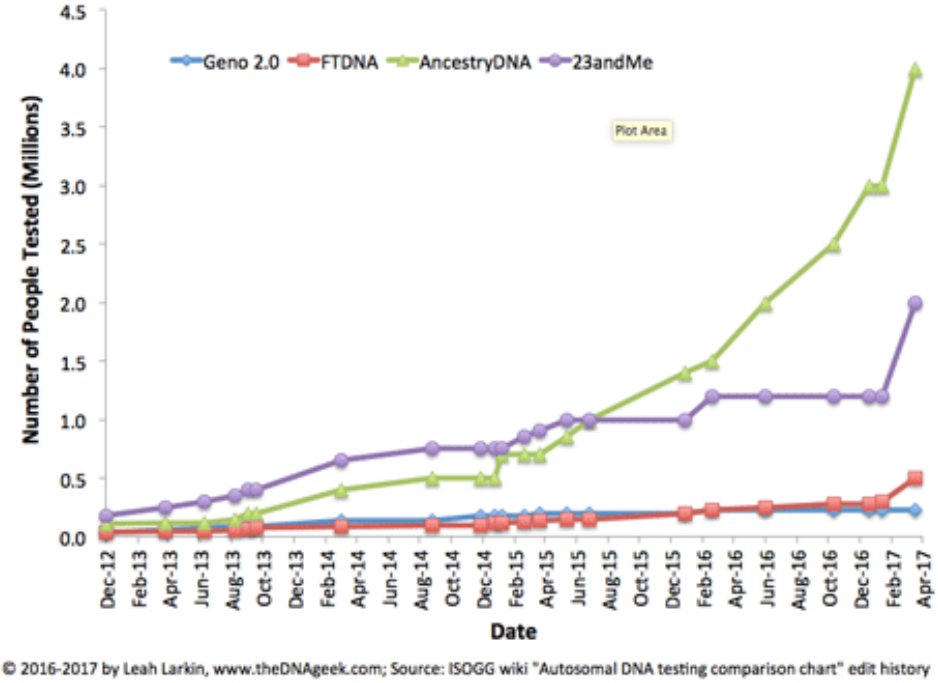


Figure 1. Recent growth in autosomal DNA testing. Data were taken from the edit history of the ISOGG wiki (http://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart). FTDNA, Family Tree DNA; Geno 2.0, National Geographic Genographic Project 2.0 and NextGeneration projects.

Journal of Genetic Genealogy 8(1):1-6, 2016

'Satiable Curiosity

Generation Gaps: A Sign of Microdeletions?

Ann Turner M.D.
DNACousins@gmail.com

'Satiable Curiosity **is a column dedicated to the proposition that genetic genealogists are an untapped resource for resolving questions about DNA behavior – how DNA changes over the course of a few or many generations and how DNA patterns are distributed around the world. Some questions are so broad that it could take decades to arrive at a conclusion, yet others are narrow enough to answer in a shorter time frame, perhaps even within a semester or two for a student research project. The results may nonetheless be of considerable genealogical utility and scientific interest, worthy of publication in a technical journal.**

Textbook genetic principles come to life when we have the opportunity to scrutinize our own data. We learn that half of our autosomal DNA comes from our father and half from our mother, and then we see it graphically illustrated, as Family Tree DNA's Chromosome Browser shows for chromosome 11:

Figure 1. Chromosome comparison showing where a tester matches his mother (orange) and father (blue). The comparison was done using the chromosome browser tool at familytreedna.com.

The microarrays ("chips") currently used by the genetic genealogy companies test about 600,000 to 700,000 single nucleotide polymorphisms (SNPs; positions known to vary) scattered over the whole genome. Each of these SNPs has two possible versions (alleles), and the probes on the chip will hunt for the presence of each allele.[1] One of the alleles found in the child can be found in the mother (the band color coded orange) and one of the alleles found in the child can be found in the father (the band color coded blue). Each chromosome is one continuous segment.

So far, so good

This level of genetic literacy lets us spot anomalies. Sometimes (rather frequently, as it turns out) there are gaps. Figure 2, clipped from a GEDmatch comparison of a parent and child, shows a break in the blue band. The yellow and green lines show SNPs where the child matches one or both of the parent's alleles, while the red lines near the center show a cluster of SNPs where a child does NOT match his parent. This leaves a gap of almost 50,000 bases.

Chr	Start Location	End Location	Centimorgans (cM)	SNPs
17	12,344	3,451,982	11.1	1,014
17	3,500,275	78,637,198	126.1	18,386

Figure 2. Chromosome comparison showing a mismatch between child and parent. The comparison was done in the one-to-one tool at GEDmatch.com with the zoom level set to 5,000 pixels.

Back to the drawing board

What is the explanation for this gap? Leaving aside the facetious suggestion that some alien DNA has infiltrated the chromosome, could the child have a bunch of mutations clustered in this location? That's exceedingly unlikely, for the mutation rate for autosomal SNPs is very low, on the order of one or two changes per 100,000,000 bases.

1 A small percentage of SNPs do have more than two known alleles, but chip technology tends to avoid those.

One possible explanation is that it is due to a limitation in the testing technology, albeit one with some interesting implications. The gap may be a microdeletion (Conrad, 2006). Microdeletions are generally defined as a loss of 1,000 to 5,000,000 bases, too small to see under a microscope with ordinary staining techniques. Recall that the chip technology looks for the *presence* of an allele. The genetic genealogy companies do not quantify the *amount* of an allele. If the base calling software sees both an A and a G for a particular SNP, it will report a heterozygous genotype of AG. If it sees only A, it will report a homozygous genotype of AA, or if it sees only G, it will report a homozygous genotype of GG.

With a deletion, one chromosome in the child will actually be missing any result for a SNP in the vicinity, and the allele from the other chromosome will be reported as homozygous. This leads to some contradictory findings: the child may be AA and the father may be GG, a "Mendelian inconsistency." According to the principles first discovered by Gregor Mendel, the father of modern genetics, the child should have at least one G.

Figure 3 shows how the actual and reported genotypes might differ in a case where the child does not match his father for a SNP. The missing allele is denoted with an x. In actuality, the child is neither *homo*zygous nor *hetero*zygous: he is *hemi*zygous. It is not possible to tell without more information whether the deletion was also present in the parent (inherited) or appeared for the first time in the child (*de novo*).

Mother		Father		Child		Type
Actual	Reported	Actual	Reported	Actual	Reported	
AA	AA	Gx	GG	Ax	AA	inherited
AA	AA	GG	GG	Ax	AA	*de novo*

Figure 3. Hypothetical example showing how a missing allele can affect reported genotypes.

A pilot study

The impetus for this column came from numerous questions about these mysterious gaps, posted on various genetic genealogy forums and mailing lists over the years. It's difficult to estimate the frequency this way. Did these queries arise from oddities and outliers, or were they perhaps the tip of the iceberg, surfacing a common phenomenon?

To approach this question, I solicited GEDmatch IDs for parent/child kits. I informed the participants that I planned to write a column, but I did not reveal the nature of my request in order to avoid ascertainment bias, where respondents might be more likely to send just the "interesting" cases.

The results were indeed intriguing. Out of a total of 86 parent/child combinations, only 11 (13%) displayed the expected number of 22 segments (one for each chromosome). The overall average was 24.7 segments, with gaps of varying sizes as shown in Figure 4.

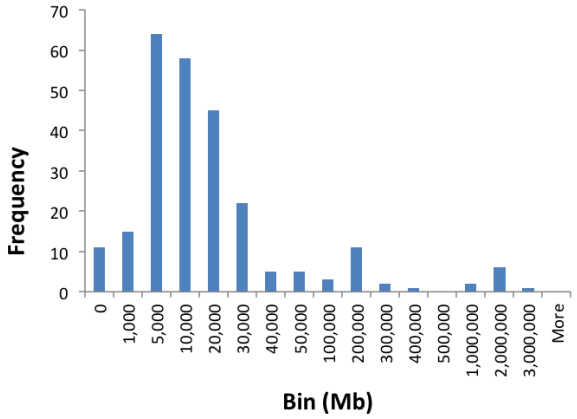


Figure 4. Distribution of 251 mismatch (gap) sizes in 86 parent/child comparisons at Gedmatch.com.

Is it real, or is it ….

That rate was much higher than I expected *a priori* (that's why we collect actual data instead of merely theorizing). It certainly convinced me that the topic merited a column, but it led inexorably to another question. How can we tell if these gaps are actually microdeletions, or if they are due to some other limitation in the testing process?

Referring back to Figure 2, there is an isolated red line toward the right edge, which does not generate a gap. There is a mismatch, but it is surrounded by matching SNPs. The genotyping process is

not perfect, and occasional miscalls are bound to occur. GEDmatch and the testing companies tolerate a mismatch or so before declaring an end to a segment, provided it is embedded in a long continuous run of matching SNPs. "Long" is not an absolute quantity, and some algorithms may be more strict than others.

Mind the gap

The mission of GEDmatch is to identify matching segments, not to analyze gaps. We are looking for mismatches that are clustered close together as a solid demonstration of microdeletions. David Pike has a suite of utilities for examining raw data, which will prove useful for digging deeper into the gaps. (This section is for those who like to get their hands dirty; others may feel free to skip ahead to the next section.)

One tool is called "Search for Discordant SNPs in Parent/Child in Raw Data Files." Discordant is synonymous with Mendelian inconsistency in this context. Figure 5 is a screen capture of some output from this utility, with columns for chromosome, reference SNP ID (rsid), position, and genotypes for the parent and child. Widely separated mismatches are included, but just eyeballing the results, it is clear that six closely spaced mismatches are found at about 112 megabases on chromosome 8.

7	rs17647441	54444505	GG	AA
7	rs2402028	115699606	TT	CC
8	rs3015792	36828215	GG	AA
8	rs1373529	93201255	TT	GG
8	rs13273246	112378777	CC	TT
8	rs1573937	112404802	GG	AA
8	rs989847	112411249	GG	TT
8	rs10094094	112417948	CC	AA
8	rs10108236	112439615	GG	TT
8	rs17520026	112488821	AA	GG
8	rs1383482	136084561	CC	TT
9	rs7863242	90780784	TT	CC
10	rs10828333	18526796	AA	GG

Figure 5. Mismatched SNPs between a parent and child. Note the cluster of mismatches on chromosome 8. The comparison was done at http://www.math.mun.ca/~dapike/FF23utils/pair-discord.php.

Supplement 1 contains a spreadsheet for automating the calculations. When data from Pike's output screen is pasted in to it, it produces summary information about the gap.

Figure 6. Summary information about the gap on chromosome 8 (Figure 5). A spreadsheet for automating these calculations is in Supplement 1.

Another level of confirmation examines the gap SNP by SNP. Referring back to Figure 3, the discrepancy is detected because the child received an A from the mother. If the mother happened to contribute a G (being homozygous GG or heterozygous AG), then the child's genotype would pass muster, masking the presence of a deletion. Figure 7 shows all of the SNPs in the gap, using David Pike's utility "Inspect a Shared DNA Segment in Two Raw Data Files."

			File1	File2	Shared
8	rs4581082	112356485	TT	TT	T
8	rs10955576	112359706	AA	AA	A
8	rs13280988	112370516	AA	AA	A
8	rs13273246	112378777	CC	TT	
8	rs1573937	112404802	GG	AA	
8	rs989847	112411249	GG	TT	
8	rs10094094	112417948	CC	AA	
8	rs13272590	112427712	AA	AA	A
8	rs6999544	112431130	TT	TT	T
8	rs10108236	112439615	GG	TT	
8	rs2606203	112454301	CC	CC	C
8	rs7005948	112455664	TT	TT	T
8	rs11991016	112463068	AA	AA	A
8	rs1366852	112463865	TT	TT	T
8	rs17520026	112488821	AA	GG	
8	rs2606199	112498181	TT	TT	T
8	rs7829881	112511904	TT	TT	T
8	rs7821076	112513833	AA	AA	A

Figure 7. All of the SNPs in the gap on chromosome 8 (Figure 5). These data were generated by http://www.math.mun.ca/~dapike/FF23utils/pair-discord.php

All the SNPs within the identified gap (and indeed for some distance beyond, not shown) are homozygous. A heterozygous result would have ruled out a microdeletion.

A man with one watch…

Do GEDmatch and the method using Pike's utilities give the same results? There's an old proverb (perhaps obsolete in the age of synchronizing our timepieces with an atomic clock) that "A man with one watch knows what time it is. A man with two watches is never sure." A spot check of some contributions to the pilot study revealed that many of the gaps in GEDmatch were not validated by Pike's

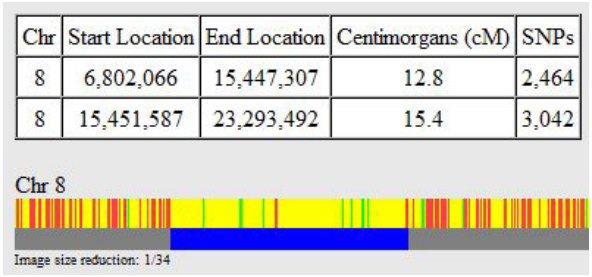
utilities. This is not to say they are false gaps – even mismatches on a single SNP could theoretically be due to a small deletion, although genotyping error rates could account for them as well. But the evidence for a microdeletion is much stronger when multiple SNPs are involved.

This dilemma is magnified by the existence of a third watch, the DNA testing companies. What do Family Tree DNA and 23andMe report using their own algorithms? A small number of cases were examined, and they showed a trend toward stitching the segments together, especially at Family Tree DNA. Closing the gaps is sensible in the framework of the big picture, but it may gloss over some informative tidbits. A larger dataset would help quantify our expectations of finding a gap. Accordingly, an online survey accompanies this column (Supplement 2). Results will be summarized in the next issue.

What's the big deal?

The preceding section was replete with obscure details about validating gaps by their content. The gaps do not challenge the parent/child relationship, so why should we bother with them, once we understand their origins? Most people don't even have a parent/child combination to check, but microdeletions can also be spotted in cousin matches. And they may make a difference in whether certain cousins are identified.

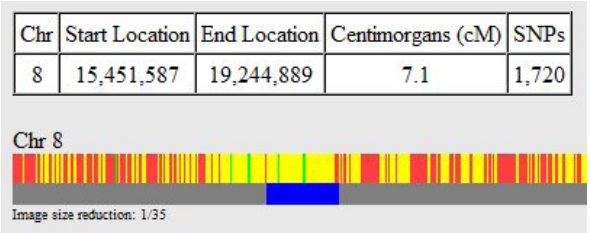
Figure 8 shows an example of a match found between a mother and a cousin, with two side-by-side segments separated by a small gap at the red bar. (The blue band appears continuous at this zoom level.) The daughter showed much the same segment boundaries, indicating that she inherited the deletion.



Chr	Start Location	End Location	Centimorgans (cM)	SNPs
8	6,802,066	15,447,307	12.8	2,464
8	15,451,587	23,293,492	15.4	3,042

Figure 8. Segment match between a mother and a cousin. The comparison was done at gedmatch.com.

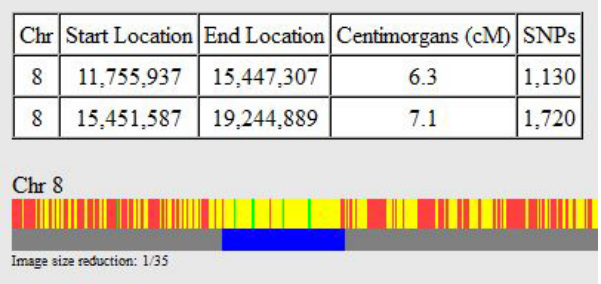
Figure 9 shows this region in a more distant cousin. Only the right-hand portion (starting at 15,451,587) registers at the default threshold of 7 cM, even though the amount of yellow and green in the left-hand side appears more prominent compared to the densely packed red bars outside of the segment.



Chr	Start Location	End Location	Centimorgans (cM)	SNPs
8	15,451,587	19,244,889	7.1	1,720

Figure 9. Segment match between the mother in Figure 8 and a more distant cousin. The comparison was done at gedmatch.com.

Figure 10 shows the match when the cM threshold is reduced to 6 cM. The portion to the left of the gap doesn't quite reach the default 7 cM threshold. In fact, the segment to the right barely squeaks by. If it had been slightly smaller, this person would not show up as a match at all.



Chr	Start Location	End Location	Centimorgans (cM)	SNPs
8	11,755,937	15,447,307	6.3	1,130
8	15,451,587	19,244,889	7.1	1,720

Figure 10. Comparison in Figure 9 done with a lower cM threshold.

Reducing the threshold at GEDmatch is often frowned upon because it can increase the number of false positive matches. However, a special dispensation may be granted when checking the extent of matching DNA next to a gap boundary.

This state of affairs is aggravating, but there may be a silver lining in the cloud. The gap may actually identify a particular lineage. Cousin 1 and cousin 2 do match each other in a portion straddling the gap (Figure 11). If a common ancestor can be identified for this group of three cousins, the deletion may tag one branch of descendants.

Chr	Start Location	End Location	Centimorgans (cM)	SNPs
8	12,637,327	19,257,994	11.0	2,791

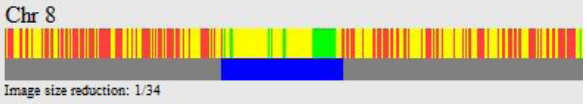


Figure 11. Match between the two cousins in Figures 8 and 9.

The five W's

Questions already abound in this column, and the end is not quite in sight. The traditional pattern of addressing the what, who, when, where, and why provides a framework.

What is the subject? A method to detect microdeletions, which cause gaps when comparing a parent and child (and cousins, too).

Who has them? Everyone, given suitable testing techniques. The Conrad (2006) study was among the first to use microarrays to identify deletions in intensively studied reference samples. "Notably, we estimate that typical individuals are hemizygous for roughly 30–50 deletions larger than 5 kb." Their samples had very dense coverage of SNPs, and the microarrays used by the genetic genealogy companies may not have enough SNPs to tag all of these.

When did they happen? Little is known about the deletion mutation rate. The deletion may have occurred in the current generation, or it may have persisted for many generations, even to the point of becoming somewhat common in the general population. The genetic genealogy community, with deep pedigrees and a propensity to test multiple family members, might provide fertile territory for a researcher seeking to determine the mutation rate. The aforementioned survey includes some questions about the gender and age of the parent. Those two factors are known to influence the rate of other types of mutations. If the number of gaps is similar for males and females and for older parents and younger parents, then that would be (very) indirect evidence that inherited mutations predominate. Gender and age would have averaged out over the generations. Conversely, differences would point to a higher mutation rate.

Where did they happen? Genealogists may wish to track the chromosomal positions as an aid in developing pedigrees, but there are also potential medical implications, depending on the location. Can we really get along without that missing DNA? Apparently so, in many cases, since we are all walking around with them. The deletions may not include genes – indeed, the likelihood is reduced by the fact that coding regions occupy only 2% or so of the genome. Even if the deletion falls in a coding region, the other copy of the gene may be sufficient, or there may alternative pathways to accomplish the task of the gene. However, there are a number of known clinical syndromes that are caused by deletions. The technical literature is voluminous; a review by Weise (2012) serves as an entry point.

Why did they happen? The most straightforward explanation is simply the lack of absolute perfection in copying DNA (slippage) or recombining it in preparation for the next generation (unequal crossing over). Recombination is usually remarkably precise, exactly trading the maternal version of a chromosomal region for the paternal version. The presence of repetitive elements in the DNA complicates matters. It's as if the enzymes lose track of where they are in the process and pick up again when they encounter something similar.

To be continued…

Results of the survey will be summarized in an anonymized and aggregated form in the next issue of JoGG. The survey does not ask for information about the size or location of the gaps to alleviate any concerns about medical privacy. Email addresses and GEDmatch IDs are optional, but if

they are provided, case studies may be used as illustrations with identifying information redacted.

The topic of microdeletions is novel territory for genetic genealogists. At the very least, this column helps explain some anomalies. Data from the survey may shed more light on whether deeper study will reap more insights.

References

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics* 38:75–81. DOI: 10.1038/ng1697

Weise A, Mrasek K, Klein E, Mulatinho M, Llerena JC Jr, Hardekopf D, Pekova S, Bhatt S, Kosyakova N, Liehr T (2012). Microdeletion and microduplication syndromes. *Journal of Histochemistry and Cytochemistry* 60:346–358. DOI: 10.1369/0022155412440001.

Conflicts of Interest

The author has had consulting agreements in the past with 23andMe, unrelated to the topic of this column.

Supplementary Information

Supplement 1: http://www.jogg.info/pages/vol8/sc/generation-gaps-Pike-template.xlsx.zip

Supplement 2: http://www.jogg.info/pages/vol8/sc/generation-gaps-survey-preview.pdf

URL for survey: https://goo.gl/forms/DyS7m79DcVmGYj182

Y-DNA Testing of a Paper Trail – The Fox Surname Project

Joseph M. Fox III and David E. Fox

Address for correspondence:
Joseph M. Fox III, 3396 Angelo Street, Lafayette, CA, USA 94549

Abstract: Combining conventional genealogical research with Y-DNA testing offers a powerful tool for confirming male lines of descent. One prominent American Fox family of Colonial Virginia had been well studied, but some relationships remained unverified. For example, was Henry Fox 2nd (1674–1750) actually the son of Henry Fox 1st (1650–1714) and Anne West? Some genealogists had denied the connection based on a lack of evidence in his grandfather's will. We analyzed Y-DNA short tandem repeat (STR) results from selected descendants of this Colonial Fox family to answer this and other outstanding questions. The DNA evidence agreed with traditional research identifying Henry Fox 2nd as the true son of Henry Fox 1st. Henry Fox 1st belonged to the R1b-L47 haplogroup, and we inferred his 37-marker STR haplotype, giving us a firm basis for comparison. DNA testing of other Fox men was able to confirm or refute proposed relationships to this family and, in the process, expand on other genealogical research efforts. Most unexpectedly, a connection to a well-researched Colonial Philadelphia Fox family was uncovered.

Introduction

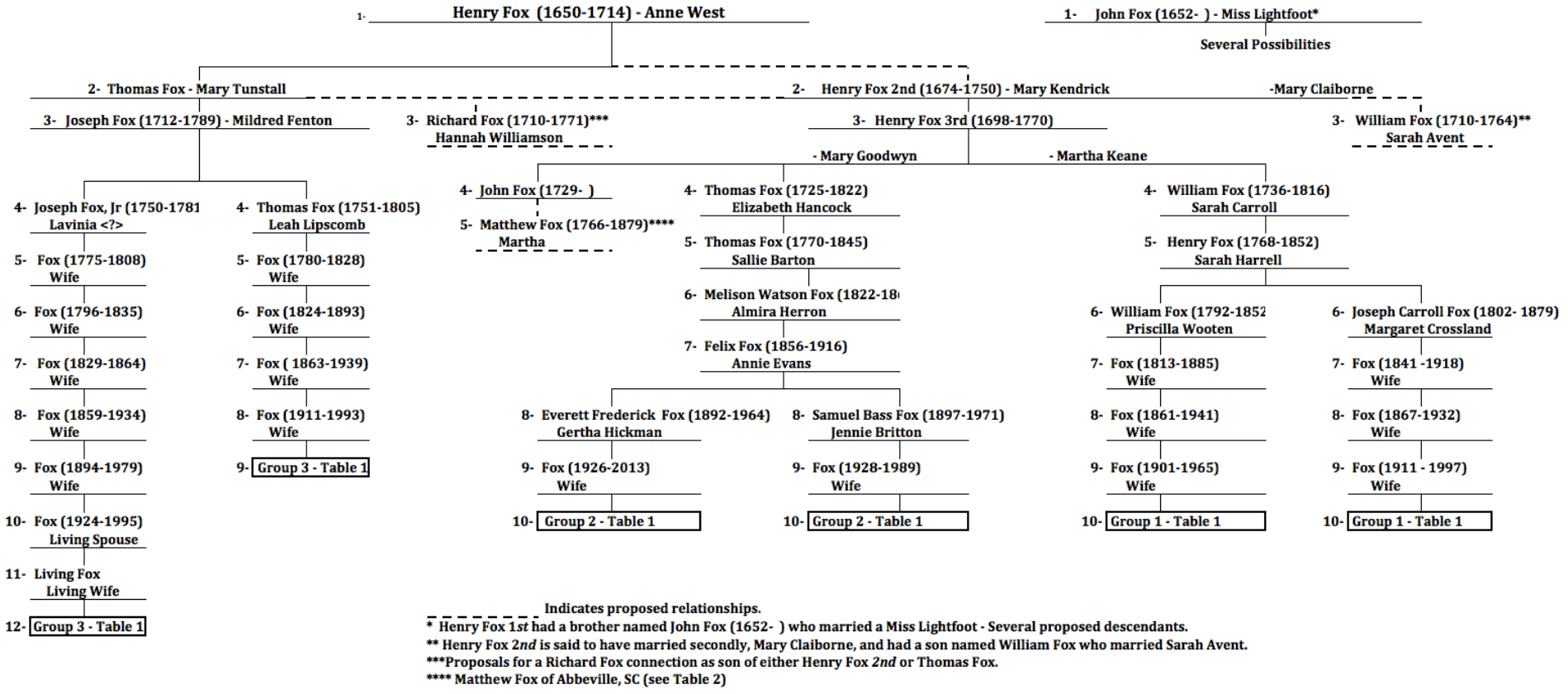
The history of the American Fox families in colonial Virginia has been exhaustively studied, particularly the ancestry and the descendants of Henry Fox 1st (1650–1714) and Anne West. Many published Fox family trees have made erroneous connections to this family, partly because this Fox line can be traced back to 1541 in Buckinghamshire, England, and partly because Anne West was the grand niece of Lord De La Warr and had royal ancestry. Probably the most thorough review of this family line was done by Joseph Steadman (1972). However, he admitted that his conclusions were often based on limited or conflicting evidence. Thus, this Fox family of Virginia offers a wonderful target for verification by genetic testing. Checking such a paper trail is by far the best way to use Y-DNA testing.

The Fox Y-DNA Surname Project was started early in 2004 with the testing of two Fox males who were thought to be about eighth cousins based on indirect, published information. When their test results matched closely, we were able to link a Colonial American Fox family from Philadelphia with their British cousins, as told in the book *Growing with America – the Fox Family of Philadelphia* (Fox, 2006). (A detailed family tree for the British family, descendants of Henrie Fox, is available from Charles Pease, Kinloch Lodge Hotel, Sleat, Isle of Skye, UK. http://kinloch-lodge.co.uk/). Using Big Y™ testing (Family Tree DNA, Houston, Texas, USA; FTDNA), we can estimate rather accurately when this link occurred. The Fox Surname Project now includes data for nearly 200 men of the Fox (or similar) surname who have tested 37 short tandem repeat (STR) markers or better, including many who traced their ancestry to Colonial Virginia, allowing us to test hypotheses that have been unresolved for more than a century.

This paper provides a number of examples from the Fox Project of the combined use of conventional genealogy with Y-DNA testing. In each case, we started with a proposed genealogical trail, identified the possible weak point, and tested descendants of several lines appropriately. This procedure has been called triangulation. The affordability and sensitivity of the 37-marker test made it the obvious choice for this study. In all cases, the 37-marker Y-DNA STR test was adequate to support a connection, and even a

Submitted 12 May 2015, Revised 7 July 2016, Accepted 13 August 2016

Figure 1. Summary of Steadman's research (1972) on the Fox lines of descent considered in this paper. Primary attention is given to Henry Fox 1st, who was born in England in about the year 1650 and married Anne West in Virginia in about 1673. Only lines that have been proposed by project members are shown. Dashed lines indicate proposed relationships.



GD refers to the genetic distance from the " ancestral" modal haplotype.
It is the same as the number of mismatches from the modal since all these are single step deviations.

12-marker test would have disproven one. In several cases, however, testing additional STR markers or SNP testing confirmed the results.

Figure 1 summarizes Steadman's research on the lines of descent considered in this paper. Primary attention is given to Henry Fox 1st, who was born in England in about the year 1650 and married Anne West in Virginia in about 1673. Steadman lists four children, Henry 2nd, John, Thomas, and Anne. However, some researchers denied that Henry 2nd (1674–1750) was their son because he was not mentioned in the will of Anne West's father. Because descendants of both Henry 2nd and Thomas have been tested, the Fox Surname Project allowed us to address this question.

John Fox, son of Henry Fox, 1st, is not shown in Figure 1 because he is not claimed as the ancestor of any Fox project members. Henry, 1st, also had a brother named John Fox (1652–) who married a Miss Lightfoot (possibly Margaret). He is shown because he has been claimed as an ancestor, since disproven by DNA testing.

Known ancestors of other Fox family groups to be discussed in this paper are:

- Richard Fox (1710–1771) who has been identified by several genealogists (but not Steadman) as the son of either Henry Fox 2nd and Mary Kendrick or Thomas Fox and Mary Tunstall;
- Matthew Fox (1766–1854), might have descended from John Fox, elder brother of Justinian. There was a John Fox on the same ship to Philadelphia who left progeny but then disappeared from the record. If this were the case, Henrie Fox would have been the common ancestor of all three groups (Fox, 2006, p. 246-248).
- William Fox (1710–1764) who married Sarah Avent, identified by Steadman as the son of Henry Fox 2nd and a second wife, Mary Claiborne.

We now report on a number of these relationships, resulting in some surprising connections to well-researched Fox family trees.

Genetic Testing of Known Henry Fox/Anne West Descendants

This paper focuses on testing a large group of STR markers on the Y chromosome. The number of repeats can be measured, and the resulting set of numbers (i.e., marker repeats) is called one's haplotype. Only men carry the Y chromosome, and the haplotype is passed almost intact from father to son. Replication errors occur frequently enough with these markers, however, that this test is useful for genealogical purposes.

The interpretation of STR testing results is governed by the laws of probabilities of rare events, and this gives a wide range of estimated generations back to a common ancestor. Estimates are based on Bayesian statistics and depend on the *a priori* probability that measured father–son marker mutation rates can be applied to the situation at hand. Results for men with different surnames should be evaluated on a more stringent basis than those where the surname is the same and, when evaluating a well researched paper trail, the *a priori* probability can be close to 1.0.

FTDNA guidelines say that if two men match on 33 or more of 37 STR markers, they are related within a genealogical time frame (Family Tree DNA, 2016). In the experience of Fox Project administrators, this approximation only holds when matches also have the Fox surname. Even then, we require members to supply ancestral information and look for common geographical locations. The 37-marker test was used because it has the highest average mutation rate of all the FTDNA panels and its affordability has made it the standard test for new project members.

The Y-DNA Haplotype of Henry Fox 1st

How do we know the haplotype of Henry Fox 1st? The answer is given in Table 1. Early on, we tested 37 markers for two well-documented Fox men (Group 1 in Table 1 and Figure 1) who descended from two different sons of another Henry Fox (1768–1852) who married Sarah Harrell, a southern USA Fox family with many living descendants (Faucette & McCain, 1971). The sons were William Fox (1791–1852) and Joseph Carroll Fox (1802–1879). The 37-marker haplotypes of these fourth cousins were identical. Another cousin (not shown) matched them on 25 of 25 markers. As a very good approximation then, this must also be the haplotype of their common ancestor, Henry Fox (1768–1852). Most genealogists accepted that this Henry Fox was the son of William Fox (1743–1816) and his wife Sarah Carroll, and the well-defined ancestry then went to William's father Henry Fox 3rd (1698–1770) and grandfather Henry Fox 2nd (1674–1750).

Secondly, we tested two second cousins (Group 2) who had identical 37-marker haplotypes to each

Table 1. STR results (37 markers) for descendants of Henry Fox and Anne West. These men belong to haplogroup R1b-L47.

Line of Descent from Henry Fox 1st	Results for Markers that Differ[1]					G.D.[1]
	DYS 458	DYS 385a,b	DYS 470	DYS 460	CDY a,b	
Group 1: Henry 1 / Henry 2 / Henry 3 / William						
Henry 1 / Henry 2 / Henry 3 / William / Henry / William / +4 gen	16	11-14	17	11	36-38	0
Henry 1 / Henry 2 / Henry 3 / William / Henry / Joseph / +4 gen	16	11-14	17	11	36-38	0
Group 2: Henry 1 / Henry 2 / Henry 3 / Thomas						
Henry 1 / Henry 2 / Henry 3 / Thomas / Thomas / Melison / Felix / Samuel / +2 gen	16	11-14	17	11	**35-38**	1
Henry 1 / Henry 2 / Henry 3 / Thomas / Thomas / Melison / Felix / Everett / +2 gen	16	11-14	17	11	**35-38**	1
Group 3: Henry 1 / Thomas						
Henry 1 / Thomas / Joseph / Joseph / +8 gen	16	11-14	17	**10**	36-38	2
Henry 1 / Thomas / Joseph / Thomas +5 gen	16	**11-11-11-14**	**18**	11	36-38	2
Group 4: Elder	**15**	11-14	**18**	11	36-38	1
Probable Ancestral Haplotype for Henry Fox 1st	16	11-14	17	11	36-38	

[1] To protect the genetic privacy of the participants, only mismatches are shown. The remaining markers were identical.

[2] G.D. refers to the genetic distance from the "ancestral" modal haplotype. It equals the number of mismatches from the modal since all these are single step deviations.

other and descended from another son of Henry Fox 3rd named Thomas Fox (1725–1822) who had married Elizabeth Hancock. Joseph Steadman (1972, pp. 54, 61) has Mary Goodwyn as Thomas' mother and Martha Keene as the mother of his half brother William. The 37-marker haplotypes of the descendants of each of these two sons of Henry Fox 3rd differed at only one marker, a multivalued, rapidly-mutating marker called CDYa,b. This five-person matchup defined the haplotype of the common ancestor, Henry Fox 3rd, except for the value at CDYa,b, and the relationship was now proven back another two generations to Henry Fox 3rd. The question still remained: Was Henry Fox 3rd (1698–1770) the son of Henry Fox 2nd (1674–1750) and the grandson of Henry Fox 1st?

In 2013 and 2014, we tested two descendants of another son of Henry Fox 1st and Anne West at 37 markers (Group 3). This son was Thomas Fox (1680–?), who married Mary Tunstall (Steadman, 1972, p. 27). These two men matched Group 1 at CDYa,b but differed from one another at DYS458, DYS385a,b, DYS460, and DYS470. On each of these four markers, however, one of them matched the first two groups, so that their genetic distance from the modal haplotype was only 2. A consensus ancestral 37-marker haplotype for all six cousins can thus be defined, and Thomas Fox and Henry Fox 2nd could indeed be considered brothers.

Finally, we have a slave descendant named Elder, who is clearly related based on his Y-DNA test results. At 37 markers, he matched our first group at all but markers DYS458 and DYS470. Since the actual connection remains unknown, further testing was deemed necessary. He and one member from each of our first two groups have been tested out to 67 markers. The two Foxes matched each other and Elder differed from them only at DYS413a,b in the last 30 markers. His results help to confirm the consensus ancestral haplotype.

Y-DNA testing had now supported that these Foxes were all one family, and we now had a good fix on the haplotype of Henry Fox 1st (Table 1). The genetic distances from the modal 37-marker haplotype were actually better than might have been expected given that Henry Fox 1st was an average of nine generations removed from each of the men tested.

In addition, Elder and a member of Group 1 have been haplogroup tested and are R1b-L47, a subclade of R1b-U106/S21. Haplogroups are defined by single nucleotide polymorphisms (SNPs) at specific sites on the Y-chromosome that mutate rarely enough to define a timeline for the history of mankind (deep ancestry). Any other Foxes who are not in this particular subclade cannot be related within the last 4,000 years (MacDonald, 2014).

The Henry Fox 2nd Controversy

Only two sons of Henry Fox 1st and Anne West were mentioned in the will of their maternal grandfather, Colonel John West, the nephew of Thomas West, Lord De La Warr. Sons John and Thomas were named, but Henry Fox 2nd was not, perhaps because Henry Fox 2nd was first in line to inherit from his father's estate (King, 1961, p. 1). Nevertheless, considerable doubt about his paternity remained in the mind of researcher Ellen Cocke (1939) and others. In 1934, Ann Woodard Fox, wife of Edward Lansing Fox, founded "The Society of the Descendants of the Hon. Henry Fox and Anne West" that claimed they were the only "approved" Virginia line; descendants of Henry Fox 2nd were not permitted to join. Both Ellen Cocke and Edward Lansing Fox were of the Thomas Fox line. Ann Woodard Fox is best known for her treatise emphasizing the royal West family connections (AW Fox 1958). She does not even mention Henry Fox 2nd.

Later researchers, including Steadman (1972, pp. 28–30) and Frances Chan (1998), felt that the overwhelming evidence was in favor of Henry Fox 2nd truly being the brother of John and Thomas. Even genealogist George Harrison Sanford King, who was the registrar of the above Society, tended to agree. Nevertheless, the seeds of doubt had been planted. The Fox Surname Project is happy now to confirm that these later researchers were correct. The close correlation between the genetic test results of Groups 1 and 2 (descendants of Henry 2nd) and Group 3 (descendants of Thomas) has shown that they were all one family.

Richard Fox (1707–1771) of Mecklenburg County, Virginia, USA

Joseph Steadman (1972, pp. 38–42) devotes several pages to various claims as to the ancestry of Col. Richard Fox who married Hannah Williamson and left many descendants. Many claims had been made that Richard was the grandson

of Henry Fox 1st and Anne West, the son of either Henry Fox 2nd or Thomas Fox. A woman even used this lineage in an application to the Society of Colonial Dames (Steadman, 1972, p. 39). Steadman disagreed and concurred with George H. S. King (1960) that he was probably the only child of a George Fox of Surry County, Virginia, though the evidence was weak. The ancestry of Col. Richard Fox remains a mystery, but Y-DNA testing is quite definite: he was not a Henry Fox/Anne West descendant. Results for a descendant of his son Jacob and a descendant of his son William match each other on 36 out of 37 markers but are a complete mismatch with the Henry Fox/Anne West descendants. In fact, they are in an entirely different haplogroup (I-L39 vs. R1b-L47).

Perhaps a clue will eventually be found from another interesting Fox Project result. Several descendants of Joaquin Fox of New Orleans, Louisiana, USA, who moved to Mexico, are obviously related to these Richard Fox descendants. One of them matches on 66 of 67 tested markers with the William Fox descendant. Given that the descendants of Richard Fox and Hannah Williamson have been well researched, this connection may well predate Richard Fox himself, even though the match is close.

Fox of Abbeville, South Carolina, USA

Henry Fox 3rd had several other sons than Thomas by his first wife, Mary Goodwyn. One of these, John Fox, was born around the year 1729. He may have participated, along with his brother Henry, in the French and Indian Wars. Steadman (1972, p. 55) has also identified him to be John Fox, a private on the payroll of Captain Andrew Miller's Company from February 1779 through May 1780 in the Revolutionary War. In this case, the Y-DNA evidence has proven Steadman to be wrong.

In December 1781, after the truce at Yorktown, Virginia, Private John Fox was captured at Pratt's Mill on Long Cane Creek by Hezekiah Williams, a Tory leader, and carried to the Cherokee Nation where he was killed. His widow, Mary (Mollie) Fox, received payment of the amount due him for service and for articles of his that were lost at Pratt's Mill. She died in 1828, and in her will she mentions a son Matthew and four daughters. Matthew Fox, born in 1766, "in Abbeville District, S.C.", enlisted at age 15 as a soldier in the Revolutionary War (Graves, 2015). He later moved to Newport, Cocke County, Tennessee, where he was living when he applied for a revolutionary war pension that confirms this information.

Matthew Fox and his wife Martha left many descendants, four of whom are in the Fox Project (Table 2, Group A). They descend from three different sons of Matthew: Anderson, William, and John S. Fox. Among 37 markers, there is only one deviation among the four of them, but they are definitely not descendants of Henry Fox and Anne West, differing by 17 or more markers (of 37) from that group. Being members of Haplogroup R1b-L1/S26, rather than Haplogroup R-L47, puts their common ancestor with the Fox/West family at thousands of years back (MacDonald, 2016). Instead, these Matthew Fox descendants are close matches at 67 or more markers with the British (Group B, Francis Fox descendants) and the American (Group C, Justinian Fox descendants) families described in detail in *Growing with America* (JM Fox, 2006, pp. 229–254). The group has now added a few more members and done more extensive testing.

A comparison of 37-marker mismatches for these three families is shown in Table 2. Josiah Fox, another of the British clan, came to America in 1793 and made a name as the designer of the *USS Constitution* (Westlake, 2003). One of his descendants was recently tested (2016), and his results are included as the third member in Table 2, Group B. He, too, had a genetic distance of 3 from the modal haplotype. The fourth member of this group was tested at 17 markers by Mark Jobling and Turi King at Leicester University in 2002. His results help define the modal values for DYS391 and DYS439. (NB: FTDNA originally read a null result at DYS439 for men in Haplogroup R1b-L1/S26 but assigned a value of 12. They changed their primer in 2014, and most of the earlier null results have now been verified as a value of 12 for the Fox group.)

Clearly, the Haplogroup R1b-L1/S26 Fox family has a higher mutation rate for their STR markers than the Henry Fox descendants (Table 1). Based on Table 2, one might expect Groups A and B to be more closely related. In fact, SNP testing has shown that Groups B and C most likely have the more recent common ancestor, confirming the genealogy trail.

With the advent of affordable Y-chromosome sequencing, we can now pinpoint the common ancestor of Groups B and C with some confidence. The Big Y test, offered by FTDNA starting in 2013,

Table 2. STR results (37 markers) for descendants of Matthew, Francis, and Justinian Fox. These men belong to haplogroup R1b-L1/S26.

Line of Descent[1]	Results for Markers that Differ[2]								G.D.[3]
	DYS 391	DYS 439	DYS 389ii	DYS 458	DYS 447	DYS 576	DYS 570	CDY a,b	
Group A: Matthew Fox (1766– ; Abbeville, SC)									
Matthew / Anderson / Matthew / James / +3 gen	11	12	29	17	25	18	17	38-38	0
Matthew / Anderson / Matthew /Henry / +2 gen	11	12	29	17	25	18	17	38-38	0
Matthew / William / +4 gen	11	12	29	17	25	18	17	38-38	0
Matthew / John / +4 gen	11	12	**30**	17	25	18	17	38-38	1
Group B: Francis Fox (1607– ; Wiltshire, Eng.)									
Henrie / Francis / Francis / George / George / +6 gen	11	12	29	17	25	**17**	17	38-38	1
Henrie / Francis / Francis / George / Joseph / +7 gen	**12**	12	29	17	25	**17**	17	38-38	2
Henrie / Francis / Francis / John / +7 gen	11	**13**	29	**16**	25	**17**	17	38-38	3
Henrie / Francis / Francis / Francis / +7 gen	11	12	29	n.a.[4]	n.a.	n.a.	n.a.	n.a.	
Group C: Justinian Fox (1673– ; Plymouth, Eng.)									
Edward/ Justinian / Joseph / Joseph / +5 gen	11	12	29	**18**	**26**	18	**16**	38-38	3
Edward/ Justinian / Joseph / Samuel / +4 gen	11	12	29	17	**26**	**19**	17	**38-39**	3
Probable Ancestral Haplotype	11	12	29	17	25	18	17	38-38	

[1] Big Y™ testing showed that Groups B and C are more closely related, and Henrie Fox is their likely common ancestor.

[2] To protect the genetic privacy of the participants, only mismatches are shown. The remaining markers were identical.

[3] G.D. refers to the genetic distance from the "ancestral" modal haplotype. It equals the number of mismatches from the modal since all these are single step deviations.

[4] n.a.: data not available because the person did not test all 37 markers

uses targeted next-generation sequencing of around 11.5–12.5 million base pairs of non-recombining Y-DNA to reveal genetic variations across the Y chromosome. One member of each of the American Groups A and C and two members of the British Group B were tested (JM Fox, 2016). All four had 20 SNPs in common downstream from L1/S26, but the British pair had one more, named A955 by YSEQ (http://yseq.net/). In addition, the two members of Group B had three singletons (private SNPs not identified in other members of Haplogroup R1b-U106/S21) between them, and the member of Group C had two singletons, indicating a close relationship. The member of Group A had seven singletons. While SNPs are random, and the number of singletons can vary, this does point to a more distant relationship.

In *Growing with America – The Fox Family of Philadelphia*, a review of genealogical evidence indicated that:

- Edward Fox was the nephew of Francis Fox (JM Fox, 2006, pp. 28, 241, 265–285), and the common ancestor of Groups B and C was Henrie Fox of Devizes, Wiltshire, England, thought to be a cousin of Sir Stephen Fox.

- Matthew Fox may have descended from an older brother of Justinian Fox who came to Philadelphia on the same ship, in which case Henrie Fox would have been the common ancestor of all three groups (J Fox, 2006, p. 246–248).

The Big Y testing results have tended to support the former conclusion, but the latter is now open to question.

The common ancestor of the British pair was George Fox, born in 1693 in Cornwall, England. A reasonable estimate of the birth date of Henrie Fox would be 1607 − 44 = 1563 (Francis, born in January 1606/07, was his 7th son). This is 130 years and three generations before George Fox, born 1693. Dr. Iain MacDonald (2016) has shown that 125 years per SNP best fits Big Y testing of Haplogroup R1b-U106/S21, of which R1b-L1/S26 is a subclade. Thus, the British pair could reasonably be expected to have experienced a SNP mutation in this time period. There is no reason to question Henrie Fox as the common ancestor of Groups B and C based on these results. If the common ancestor of the American Foxes was Edward Fox of Plymouth, then their common ancestry back to Henrie Fox would have been two generations and about 80 years, during which time they might well not have experienced a SNP mutation. The high number of singletons for the Group A member, however, tends to suggest an earlier common ancestor.

Another interpretation is that Group A has a direct connection to Sir Stephen Fox. Burke's *Landed Gentry* says that Francis Fox was "stated to be of the same family as the celebrated Sir Stephen Fox, ancestor of the Earls of Ilchester and the Lords Holland" (Burke & Burke, 1847, p 441), and the Francis Fox family is permitted to use his coat of arms.

In his book on the Fox family, James Wallace Fox (1917, p. 8) relates several tales of how Sir Stephen's grandson, the politician Charles James Fox (1749–1806), corresponded with and sent gifts of jewelry to several Fox relatives of his in Virginia. This jewelry ended up in the hands of another Charles James Fox, a bachelor who was said to be the son of John and Grace Fox. Unfortunately, they all ended up in the possession of relatives named Moody or Montague and were lost or stolen. Could this actually be the Matthew Fox line that Steadman incorrectly identified as Henry Fox/Anne West descendants? The Big Y results suggest so. There are known descendants of Sir Stephen Fox living in England and, hopefully, further testing will tell the tale.

William Fox of Loudoun County, Virginia

Another Virginia Fox family that has often been confused with the Henry Fox/ Anne West family is that of William Fox, Sr., born about 1710 in Loudoun County, Virginia. The descendants of his son, William Fox, Jr., were well covered in a book by Nellie Fox Adams (Adams & Walton, 1998). John Fox, the author of *Little Shepherd of Kingdom Come*, the first American novel to sell 1,000,000 copies, was from this line.

This is also the family line of James Wallace Fox who wrote *Fox Family* (1917). At the end of this book, he mentions James Fox who married Mary Bartleson at Swede's Church in Philadelphia on 1 September 1758, but fails to connect him to William Fox, Sr. There is now good Y-DNA evidence that James Fox and William Fox, Jr., were brothers and

the sons of William Fox, Sr., and his wife Elizabeth.

Two project members (Table 3, Group 1) descend from William Fox, Jr., and one member (Group 2) descends from the James Fox who married Mary Bartleson. They are exact matches at 37 markers, confirming the relationship. The comparison is carried out to 67 markers in Table 3 to accommodate test results from others who appear to be related.

Joseph Steadman (1972, p. 19) has James Fox and his son Bartleson Fox as possible third- and fourth-generation descendants of John Fox, brother of Henry Fox 1st. As shown in Figure 1, John Fox was born about 1652 and reportedly married a Miss Lightfoot. Steadman guessed wrong. The two groups have a genetic distance of 23 based on 67 markers and this family is a predicted member of haplogroup R1b-L21. R1b-L21 is a subclade of R1b-P312 and any connection with the Henry Fox/Anne West line goes back at least 5,000 years (YFull Tree, 2016).

Group 3 in Table 3 includes two members whose lines of descent are not yet confirmed but who are undoubtedly related to the William Fox, Sr., family. They each match the Group 2 descendant on 66 of 67 markers. The John Fox (1780–1852) descendant is positive for the S1051 SNP and is a member of the R1b-S1051 Project, as is the first member of Group 1. R1b-S1051 is a subclade of R1b-L21 that may have originated in what is now Scotland. Both men also matched each other in the defunct Relative Genetics database of Ancestry.com (Lehi, Utah, USA).

The John Fox descendant originally proposed the following ancestry (personal communication): Henry Fox (Anne West) → Thomas Fox (Mary Tunstall) → Joseph Fox (Mildred Fenton) → Thomas Fox (Leah Lipscomb) → John Fox, Jr. ~1780 VA and KY (Elizabeth Hoffman). However, the family tradition was wrong, given his haplogroup. Based on the research of Kevin Daniel, who has an online Fox family tree (Daniel, 2001), and Jane Fox Wheldon, who has researched the Bartleson Fox line (personal communication), both John Fox and Enos Fox are thought to be later descendants of James Fox by his second wife.

Another pair of men in this lineage serve as an example of a match that requires further study. Two descendants of Hugh Fox, born about 1745 in Virginia, match the William Fox, Sr., descendants at 32 and 34 of 37 markers, respectively (Table 3). The 32 for 37 marker match is also a 60 for 67 marker match. A third Hugh Fox descendant elected to test only 12 markers but confirms the mismatch at DYS389ii.

These less-close Y-DNA test results indicate that a possible long-range family connection may exist within a genealogical time frame. Further testing is recommended, and the S1051 SNP is an obvious choice for either joining or separating these two family lines. Current thinking is that all these Foxes may have come down to Virginia from Philadelphia or New Jersey, which may explain James' marriage back in Philadelphia.

William Fox (1710–1764) Who Married Sarah Avent

Perhaps the most interesting of the erroneous Fox relationships, because it had been so abundantly documented, is that of two descendants of William Fox of Virginia (b: 1710) and his wife Sarah Avent. In Shirley Faucette's (1972, pp. 119–124) comparison of the genealogists Steadman and Robinson, both have this William Fox as the son of Henry Fox 2nd. Steadman (1972, p. 28) comments:

> "The said William Fox doubtless was that one who settled in Brunswick County (Virginia), being named as the son of Henry Fox 2nd and Mary Claiborne. He married Sarah Avent who was a granddaughter of William Gooch and his wife Ursula Claiborne. — See Joseph Emery Avent's 'The Avents and Their Kin of Avent Ferry, Chatham County, North Carolina'."

We now have conclusive Y-DNA evidence that William was not the son of Henry Fox 2nd. With the help of Fox researcher Donald Fletcher, known descendants of two sons of William Fox and Sarah Avent, John and Thomas, were located. They have been tested on 37 markers and they differ at DYS385a,b, CDYa,b, and DYA442 (11-13, 37-37, and 16, respectively, for the descendant of John; 11-14, 37-39, and 14 for the son of Thomas). This large a difference is unusual but not unexpected for two men whose common ancestor is seven generations removed. A first cousin of the Thomas Fox descendant has been tested on 12 markers and is an exact match on the first 12, which includes DYS385a,b.

These men differ, however, on 17 or more out of 37 markers from our Henry Fox/Anne West descendants.

Table 3. STR results (67 markers) for descendants of William Fox, Sr. and Hugh Fox. These men belong to haplogroup R1b-L21.

Line of Descent	Results for Markers that Differ[1]								G.D.[2]
	DYS 389ii	DYS 448	DYS 456	DYS 576	CDY a,b	DYS 578	DYS 557	DYS 444	
Group 1: William / William									
William / William / James / James / +4 gen	31	19	16	18	35-38	n.a.[3]	n.a.	12	0
William / William / James / Rueben / +3 gen	31	19	16	18	35-38	n.a.	n.a.	n.a.	
Group 2: William / James									
William / James / Bartleson / +4 gen	31	19	16	18	35-38	8	16	12	0
Group 3: John and Enos Fox									
John Fox (b1780 VA, d1852 KY) / +5 gen	31	19	16	18	35-38	8	**15**	12	1
Enos Fox (b1814 KY, d1897 IA) / +4 gen	31	**20**	16	18	35-38	8	16	12	1
Probable Ancestral Haplotype for William Sr.	31	19	16	18	35-38	8	16	12	
Group 4: Hugh Fox Descendants									
Hugh / Hugh / James / +6 gen	**30**	19	**17**	**19**	**36-37**	**9**	16	**13**	
Hugh / Moses / Hugh / +5 gen	**30**	19	**17**	18	**36-38**	n.a.	n.a.	n.a.	

[1] To protect the genetic privacy of the participants, only mismatches are shown. The remaining markers were identical.

[2] G.D. refers to the genetic distance from the "ancestral" modal haplotype. It equals the number of mismatches from the modal since all these are single step deviations.

[3] n.a.: data not available because the person did not test all 67 markers.

The John Fox descendant has been tested on 67 markers and differs from them on 24 markers. In addition he has 12 repeats at stable marker DYS492, and the Henry Fox/Anne West descendants have 13 repeats. This result points to haplogroup R1b-P312, whereas the Henry Fox/Anne West descendants are in the R1b-L47 subclade of R1b-U106. This would put their common ancestor back some 5,000 years (YFull Tree, 2016). The published information is wrong.

We are not even certain whom Henry Fox 2nd married. Shirley Faucette (1972, p. 121) states that, "Some sources list both wives, others show only one but vary as to whether it was Mary Kendrick or Mary Claiborne." It is quite possible that the Henry Fox who married Mary Claiborne was a different person than Henry Fox 2nd, son of Henry Fox 1st and Anne West.

Interestingly enough, William Fox and Sarah Avent were the grandparents of Sarah Harrell, the spouse of Henry Fox (1768–1852) of Webster County, Mississippi, ancestor of Group 1 of the Henry Fox/Anne West descendants. One of Henry Fox/Sarah Harrell descendants, Frances Cooke Chan (personal communication), writes, "I don't think anyone in our family ever felt that they (Sarah Harrell's grandparents) necessarily were in this Fox family, just that they had the same name and might have been relatives."

Andrew Fox (1749–1819) of Virginia and Tennessee

A classic example of how erroneous family trees gain credence is the tale of Andrew Fox, who first appeared in Culpeper County, Virginia, in 1772 and then showed up in Greene County, Tennessee, in 1786. Three of his descendants have been tested at 37 markers. None of them match our Henry Fox/Anne West descendants, and there is a genetic distance of 21 to 24 for the 37 markers.

Someone, however, had reported a connection to Henry Fox and Anne West via Henry Fox 2nd and Mary Kendrick, and then via a son named Jacob, a connection that managed to get into the files at the Family History Library (Salt Lake City, Utah, USA). Once there, the relationship was considered documented by many others and published on various internet sites. One classic example is the Germanna Research site (Blankenbaker, 2008), which questions contrary evidence published by a researcher named John Fox (2004) and says Andrew may have been of German origin, yet still uses the Family History Library tree. The researcher John Fox had suggested that Andrew Fox was the son of a pauper named Anne Fox and came as an indentured prisoner to Culpeper, Virginia, in 1772 from Rutland, England.

James Fox, in his book *Tracking Andrew Fox* (2012), concludes that John Fox was correct. He says that Andrew Fox was indeed the illegitimate son of Anne Fox but thrived in America, serving in the Revolutionary War; marrying Sarah Render of Culpeper, Virginia; and acquiring 300 acres of property in Tennessee. The evidence is circumstantial, but Y-DNA testing tends to confirm this version over the others. Andrew Fox was definitely not a Henry Fox/Anne West descendant, and his father may not have been a Fox.

There is evidence for a possible non-Fox connection. A comparison at 37 markers between our three Andrew Fox descendants and a man with another surname who traces back to Scotland in 1898, is shown in Table 4. Significantly, the non-Fox matches the ancestral value for all markers except CDYa,b. This is a close match indeed and tends to confirm the Andrew Fox story, but further testing is required before we can confidently say that this is the connection. The non-Fox descendant has been tested out to 111 markers and his haplogroup assignment has been confirmed as R1b-DF13 (a subclade of R1b-L21) by SNP testing, a result possibly indicative of an ancient Scots/Irish ancestry. If one of our Andrew Fox descendants were to upgrade, the results might well solidify the connection.

Other Virginia Fox Families

More than a dozen other Fox Project members erroneously thought they might be descendants of Henry Fox and Anne West. This list includes a descendant of William Eires Fox (b. 1758 in Virginia), a descendant of Allen Fox (b. 1760 in North Carolina), a descendant of John Fox (b. ca. 1705–15 in Essex County, Virginia), two descendants of John B. Fox (b. 1745 in Orange County, Virginia) and his wife Ann Barber, and two descendants of William Fox (b. 1836 in Warwick County, Virginia), whose parents were William Fox and Nancy Stacy.

Table 4. STR results (37 markers) for descendants of Andrew Fox of Virginia and Tennessee. The non-Fox belongs to haplogroup R1b-DF13, a subclade of R1b-L21.

Line of Descent	Results for Markers that Differ[1]				G.D.[2]
	DYS 449	DYS 576	DYS 570	CDY a,b	
Andrew / Jacob / Matthias / +5 gen	28	**20**	**19**	36-38	2
Andrew / Jacob / Joseph / +4 gen	28	**18**	18	**36-40**	2
Andrew / Jesse / +5 gen	**29**	19	18	36-38	1
Non-Fox (Scotland 1898)	28	19	18	**36-37**	1
Probable Ancestral Haplotype	28	19	18	36-38	

[1] To protect the genetic privacy of the participants, only mismatches are shown. The remaining markers were identical.

[2] G.D. refers to the genetic distance from the " ancestral" modal haplotype. It equals the number of mismatches from the modal since all these are single step deviations. Mutations in multicopy marker CDYa,b are considered a single step.

Conclusions and Recommendations

As we have seen, 37 markers can be sufficient to deny a relationship and can confirm one when a paper trail is available and multiple descendants are tested. When there is no paper trail and the surname differs, additional DNA testing is required.

There must be hundreds of erroneous Fox genealogies posted on the internet that rely on sources mentioned here. This paper cannot resolve all these problems but perhaps makes a good start. As the public comes to realize the power of genetic surname testing, they will hopefully correct most of these errors. Those whose connection to Henry Fox and Anne West was disproven have already defined new goals for their research. Those whose connection was proven can rejoice that a contentious issue has finally been settled.

Many challenges remain, and perhaps this paper will spur more people to help resolve them. The many Virginia John Foxes remain something of a mystery. As mentioned previously, Henry Fox 1st had a brother named John Fox who married a Miss Lightfoot. Hopefully, there is a direct male descendant of this line we can locate and test. The English ancestry of Henry Fox 1st is an important unanswered question, as is the relationship of the Haplogroup R1b-L1/S26 Foxes to Sir Stephen Fox.

Ann Woodard Fox took the ancestry of Henry Fox 1st back to England, and Joseph E. Steadman (1972, pp. 3–11) later made a comprehensive review of what is known about the British ancestry of Henry Fox 1st. He was the son of John Fox, a sea captain who also settled in Virginia in 1661, and this line has been tentatively traced back to Henry Fox (1521) who married a Hawes of Missenden and possibly to a William Fox (1497–1559) of Missenden, Buckinghamshire, who lived at Stewkley Manor (J William Fox, 2004).

A William Vaux, descended from a Norman Invader named Robert de Vaux, is known to have inherited Stewkley Manor by marriage in 1424. Some researchers question a change from Vaux to Fox but, if a Fox/Vaux connection could be substantiated, this would carry the line back to 1066.

Acknowledgments

The authors thank three anonymous reviewers for their valuable comments, which helped us to improve the manuscript.

Conflicts of Interest

The authors declared no conflicts of interest. Joe Fox is a retired Process Design Manager at Bechtel, Inc., San Francisco, CA, USA, and the administrator of the Fox Y-DNA Surname Project at FTDNA.

References

Adams NF, Walton BF (1998) Fox Cousins by the Dozens. Higginson Book Co., Salem, MA. (reprint of 1976 book)

Blankenbaker J (2008) http://wc.rootsweb.ancestry.com/cgi-bin/igm.cgi?op=GET&db=germanna&id=P18346. Accessed Aug 2016.

Burke J, Burke JB (1847) A genealogical and heraldric dictionary of the landed gentry of Great Britain and Ireland, Vol. 1. Henry Colburn, London.

Chan FC (1998) Ancestors of Anselm Cooke. Langford Pub., Hervey Bay, Queensland.

Cocke EM (1939) Some Fox trails in old Virginia: John Fox of King William County, ancestors, descendants, near kin. Dietz Press, Petersburg, VA.

Daniel KWQ (2001) Kevin Daniel's Genealogy. http://wc.rootsweb.ancestry.com/cgi-bin/igm.cgi?op=GET&db=kwdaniel&id=I81. Accessed 24 Jun 2013.

Family Tree DNA (2016) Family Tree DNA Learning Center: Y-DNA – Matches Page. https://www.familytreedna.com/learn/user-guide/y-dna-myftdna/y-matches-page/. Accessed Aug 2016.

Faucette S (1972) Steadman vs. Robinson. In: Steadman JE Sr (1972) Ancestry of the Fox family of Richland and Lexington Counties, South Carolina. https://dcms.lds.org/delivery/DeliveryManagerServlet?dps_pid=IE1044049

Faucette S, McCain WD (1971) An outline of four generations of the family of Henry Fox (1768–1852) and his wife, Sarah Harrell Fox (1772–1848), of South Carolina and Mississippi. Self published by WD McCain, Hattiesburg, MS.

Fox AW (1958) The noble lineage of the Delaware–West family of Virginia through Col. John West his sons and his daughters including Ann West who married Henry Fox. M McNeill Ayres, ed. Seebode Publishers, Memphis, TN.

Fox John (2004) The Andrew Fox family tree. http://wc.rootsweb.ancestry.com/cgi-bin/igm.cgi?op=GET&db=jf5&id=I0925. Accessed 29 Nov 2014.

Fox James (2012) Tracking Andrew Fox. Create Space Publishing, Charleston, SC.

Fox JM (2006) Growing with America – The Fox Family of Philadelphia. Xlibris, Bloomington, IN.

Fox JM (2016) STR vs SNP Big Y 29. https://groups.yahoo.com/neo/groups/R1b1c_U106-S21/files/%20Admin%20files/

Fox J Wallace (1917) Fox Family. Reprint from the October issue of William and Mary Quarterly. Whittet & Shepperson Printers, Richmond, VA.

Fox J William (2004) Fox/Vaux Lineage. http://growingwithamerica.weebly.com/uploads/5/9/3/3/59339633/fox-vaux_genealogy_by_jwfox.pdf. Accessed Aug 2016.

King GHS (1960) Letter to Mrs. Vivian T. Rousseau, October 6, 1960. In the George Harrison Sanford King (1914–1985) papers, Virginia Historical Society, Richmond, VA.

King GHS (1961) Letter to Dr. M. Harris, April 27, 1961. In the George Harrison Sanford King (1914–1985) papers, Virginia Historical Society, Richmond, VA.

Graves W (2015) Pension Application of Matthew Fox 3279, transcribed by Will Graves. http://revwarapps.org/r3729.pdf. Accessed Aug 2016.

MacDonald I (2014) Haplogroup U-106/S21 Family Tree, pdf file updated November 24, 2014. https://groups.yahoo.com/neo/groups/R1b1c_U106-S21/files/%20Admin%20files/. Accessed Aug 2016.

MacDonald I (2016) U106 Overview. https://groups.yahoo.com/neo/groups/R1b1c_U106-S21/files/%20Age%20Analysis/. Accessed Jul 2016.

Steadman JE Sr (1972) Ancestry of the Fox Family-of Richland and Lexington Counties, South Carolina. https://familysearch.org/search/catalog/271092?availability=Family%20History%20Library.

Westlake M (2003) Josiah Fox 1763–1847. Xlibris, Bloomington, IN.

YFull Tree (2016) https://www.yfull.com/tree/R1b/.

Evidence of early gene flow between Ashkenazi Jews and non-Jewish Europeans in mitochondrial DNA haplogroup H7

Doron Yacobi & Felice L Bedford, Ph.D.

Address for correspondence:
University of Arizona, P.O. Box 210068, Tucson, AZ 85721, USA

Abstract

To investigate European introgression into Ashkenazi Jewry, the European-dominant haplogroup H mitochondrial DNA was examined. The results provided genetic evidence that gene flow between Jewish and non-Jewish populations occurred early in Jewish settlement in Europe with isolation of the groups thereafter. We targeted branch H7 and found three Ashkenazi Jewish clades, two that were not previously recognized as Jewish (H7e, H7c2) and one newly identified group (tentatively H7j) characterized by 1700C and 152C transitions. A total of 100 new complete mitochondrial DNA sequences (mitogenomes) are reported, including the largest collection of H7e to date. H7e is a deeply nested clade with several subclades; more than 85% of the carriers had Ashkenazi maternal ancestry from such diverse areas as Germany and Austria in Western Europe, Poland, and the Baltic states in Central Europe, and Moldova, Ukraine and Belarus in Eastern Europe. Between 10% and 15% of the carriers had European non-Jewish ancestry which, strikingly, showed the greatest number of mutational differences from ancestral H7e. Moreover, there was no overlap with the Jewish-affiliated sequences other than at the ancestral node. Earlier research proposing early mixing followed by isolation has relied on less direct inferences. The smaller groups of H7c2 and H7j were exclusively Ashkenazi Jewish, with interesting sequence patterns. H7c2 consisted of a number of non-nested sister branches, reflecting recent expansion in a large population, while H7j showed a possible in-progress vanishing of the ancestral group, well on its way to mothering an orphan node. The severe bottleneck and subsequent population explosion in the Ashkenazim provide a unique opportunity to view haplogroups in all states of evolution and provide a window into the Mediterranean–Hellenistic world of antiquity.

Introduction

Over the last decade, evidence has accumulated that the genetic make-up of Ashkenazi Jewry is a combination of Levantine and European sources. Analyses of autosomal genes, reflecting a combination of paternal and maternal inheritance, have indicated a significant degree of European admixture among Ashkenazi Jews as well as a close relationship between most contemporary Jews and non-Jewish populations from the Levant (Atzmon et al., 2010; Behar et al., 2010). The source of the European contribution may come from the maternal line. Costa and colleagues (2013) argued that the majority of the Ashkenazi mitochondrial haplogroups, which are inherited only from the mother, were present in Europe long before the arrival of Jews. However, Behar and colleagues (2006) suggested that these same maternal haplogroups most likely originated in the Levant alongside paternally inherited Y chromosomes of Levantine origin (Atzmon et al., 2010; Ostrer & Skorecki, 2013).

When haplogroups have a notable presence in both the Near East and Europe, determining their geographic origins can be challenging and lead to differing interpretations. An example involves T2e, a haplogroup that harbors a couple of unique Jewish clades. Bedford (2012) reported prevalence of

T2e in Italy, Egypt, and parts of Saudi Arabia and favored a Near Eastern rather than European origin of the mutations that define T2e but left open the possibility that either locale could be the origin or recipient of migration. On the other hand, Pala et al. (2012), using similar geographic incidences, concluded that T2e's origin was European.

In principle, estimates of when mutations emerged can help resolve where they emerged. In practice, however, standard deviations of time estimates can extend across greater than a thousand years, and time estimates themselves can differ by an order of magnitude depending on the estimated mutation rate. In research on Jewish groups, we (Bedford et al., 2013; Bedford & Yacobi, 2014) reported on a Bulgarian Sephardic founding lineage (T2e1b), originally identified by Behar, which we found among both Ashkenazi and Sephardic Jews from diverse regions. Full genomic sequencing found much coding-region variability, with several haplotypes. Coalescence time for the sequences using a common mutation-rate estimate suggested that the shared mutation (9181G) predated the split between the Jewish groups and therefore likely arose in the Levant. However, a different , also justifiable mutation rate suggested the origin was much more recent, implicating geneflow in Europe after the split as the source as of the mutation common to both Sephardic and Ashkenazi populations.

Difficulty in distinguishing between Levantine and European sources for Ashkenazi mitochondrial haplogroups is further muddled by an often overlooked historical fact: that the boundaries of Europe and the Levant are a relatively recent historical construct dating back to the Arab conquest in the 7th century CE.

To further investigate the role of European maternal admixture into the Ashkenazi gene pool, we took a different approach than previous investigations. Rather than surveying a large number of haplogroups with ambiguous geographic origins, we conducted a detailed investigation into a haplogroup that is overwhelmingly European (e.g., Brotherton et al., 2013) yet still found among modern Ashkenazi Jews. Haplogroup H is the dominant European mtDNA haplogroup. Its numerical success nears half the population in some countries, making it the most common haplogroup in Europe. Among Ashkenazi Jews, 23% have haplogroup H (Costa et al., 2013), yet despite being a "major" Ashkenazi haplogroup, it is often overlooked. When examining Ashkenazi H mitogenomes, Costa and colleagues found that most of them nest within west/central European subclades, with closely matching sequences in Eastern Europe. As such, haplogroup H's general European dominance may illuminate issues of introgression of European DNA into the Ashkenazi gene pool. Does haplogroup H reflect recent unions of non-Jewish women and Ashkenazi men, or does it point to events of more distant interest?

We focused on H7. While other choices were possible, we selected H7 as an understudied clade within haplogroup H that our pilot study suggested had an unexpected notable presence among the Ashkenazim. Finally, we also delved into Mediterranean and Jewish history to place the genetic results within their correct historical framework. A consideration of relevant Mediterranean and Jewish history is given in Appendix A. The combination of genetic results and accepted history may lead to a greater understanding of Jewish maternal lineages.

Materials and Methods

To identify Ashkenazi clusters within haplogroup H7, we initially selected two individuals with self-described Ashkenazi Jewish maternal lineages belonging to two different subclades of H7 from the customer base at Family Tree DNA (FTDNA; Houston, Texas, USA). FTDNA offers genetic testing services direct to individuals and has one of the largest databases in the world of individuals who have had their full mitochondrial genomes sequenced, including many with European and Ashkenazi Jewish roots. The data from FTDNA customers is increasingly being used as a scientific resource (Bedford, 2012; Bedford et al., 2013; Behar et al., 2012; Pike, 2006; Pike et al., 2010).

These two sequences were used as "kernels", or seeds, to search the FTDNA database for other full

mitochondrial sequences that differed by 0–3 mutations, as in our previous study (Bedford et al., 2013). These people were contacted by email and invited to be part of the research study. They were asked about 1) the additional mutations they carried in their mtDNA, 2) who their matches were within 0–3 genetic differences, and 3) their deep maternal ancestry. In this manner, a large number of different haplotypes belonging to both H7 subclades was identified, and a robust picture of all members of these Ashkenazi Jewish clusters was assembled.

Thereafter, the database of the H7 mtDNA genome project ("H7 MtGenome"), co-administered by one of us (Yacobi), was mined for additional sequences not uncovered by the above procedure. Within the H7 MtGenome project, 229 participants had tested their full mitochondrial genome at the time of this study. The H7 MtGenome project is open to anyone who has tested their mtDNA full genomic sequence with FTDNA and belongs to H7 or one of its subclades (https://www.familytreedna.com/public/mtdna_h7/). All members who were not contacted initially and whose data showed they belonged to one of the groups of interest (the two identified Jewish clades and any cluster which suggested Jewish presence) were also issued invitations to participate in the study and questioned as above.

In addition, for each Ashkenazi cluster found, a sister cluster was sought for comparison among project members without regard to ethnicity. Sister clusters were defined as two distinct branches deriving from the same mother node in the tree. Sequences will also be deposited in GenBank (see Supplementary Table 1).

We decided to use relative time origins, where appropriate, rather than ambiguous absolute time estimates.

Results

Three branches with a notable Jewish constituency were identified within haplogroup H7, for a total of 89 sequences. Two of these branches, H7c2 and H7e, have been previously identified but not previously connected to Ashkenazi Jewish roots. The third branch is newly reported here; it is defined by a nucleotide transition from T to C at position 1700 in the coding region and by two additional mutations (152C, 573.1C), and thus was not identifiable from inspection of the first control region alone. We tentatively label this clade H7j, following standard mtDNA nomenclature (Phylotree Build 17; Van Oven, 2015). The three branches likely represent three different maternal founders. In addition, two sister clades were identified for H7c2 among the project's participants, namely H7c1 and H7c3, both documented branches of H7c. We did not find any sister clades to H7e or H7j in our data set. An overview of the five branches in relation to the H7 ancestral node is shown in Figure 1.

H7J

A total of 14 individuals belonging to newly identified H7j were found. Of these, nine agreed to participate. All nine participants reported Ashkenazi Jewish ancestry on their direct maternal line, with one noting additional possible ancient Sephardic Jewish roots. A notable pattern was observed in this small clade in which the most frequent sequence was not ancestral H7j, but rather a descendant branch (see Figure 1, bottom branch). There is no known positive selection pressure because its single change in the coding region (T11137C) is a synonymous mutation. The success of this branch within H7j may instead be due to random drift during the population explosion following the severe Ashkenazi bottleneck. We may be witnessing the in-progress disappearance of the mother node of H7j, which is becoming less prevalent than its daughter node, presumably an intermediate step before being lost entirely to history and producing breaks in the phylogenetic tree.

H7c2 and sister clades H7c1 and H7c3

A total of 25 people were found in H7c2, 17 of which responded to the invitation. All 17 reported Ashkenazi Jewish ancestry on the direct maternal line. We do not think this reflects sampling bias because public information available on individuals who did not respond pointed to Ashkenazi

Jewish ancestry as well. H7c2 consisted of individuals from regions of Austria, Hungary, Poland, Romania, and the Pale of Settlement.

Of the 25 individuals confirmed as belonging to H7c2, a large majority (20) belonged to the ancestral cluster (A13959T). The remaining five each had a unique haplotype. This is consistent with recent expansion in a large population, large enough for several branches to emerge contemporaneously. The deepest nesting was separated by two mutations from the ancestral H7c2, belonging to an individual of Hungarian Jewish ancestry (see Figure 1).

In contrast, the sister clade H7c1 (previously estimated to be over 3,000 years old; Behar et al., 2012) had a wider geographic distribution than Ashkenazi dominated locales, with our participants reporting ancestry from Egypt, Asia Minor, Italy, Germany, the British Isles, and the Ukraine. H7c1 is also found among the Druze of Israel (Shlush et al., 2008). One of our participants reported Sephardic Jewish ancestry, and the remaining participants denied any Ashkenazi Jewish ancestry. The current distribution of H7c1 may reflect population movements around the Mediterranean during and subsequent to the Roman era.

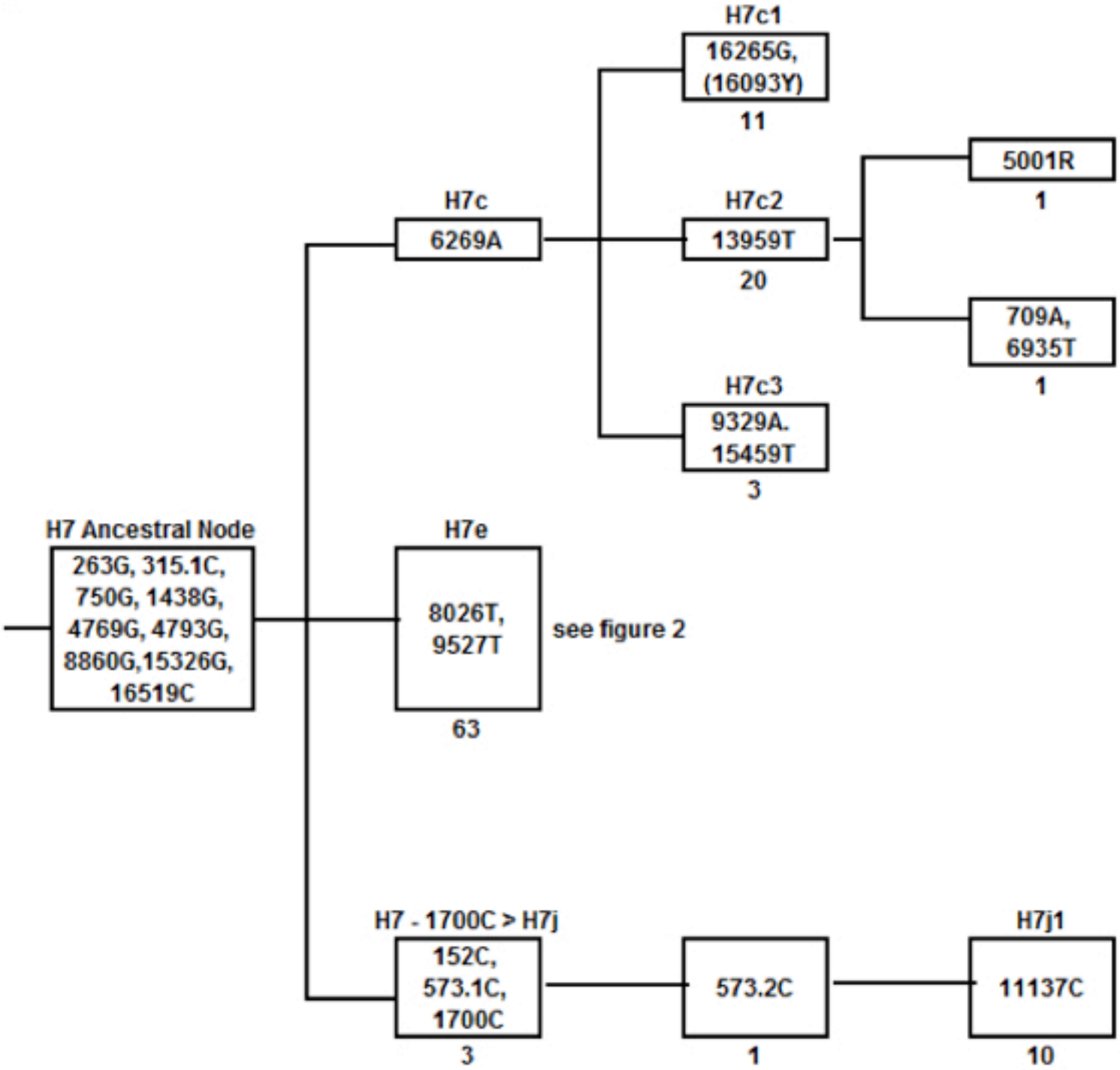
The second sister clade H7c3 (estimated by previous researchers to be 2440 years old) was distributed mainly in Northern and Eastern Europe with ancestry reported from Finland, Sweden, Russia, and Poland. As with H7c1, no individuals with Ashkenazi Jewish ancestry were reported despite the haplogroup being found in some of the areas heavily populated by Ashkenazi Jews, such as Galicia in Poland.

The Ashkenazi Jewish H7c2 appears to be a younger clade than sister H7c1 with one fewer mutation separating it from the mother haplogroup H7c and less rich nesting structure. H7c2 has been dated previously to 1,735 YBP (Behar et al. 2012), younger than the 3000+ YBP estimate for H7c1 and 2400+ YBP for H7c3. The relatively young cluster of H7c2, found here only in Ashkenazim (although among multiple diverse communities), favors a local European emergence in early Ashkenazi settlement predating their geographic dispersal. In view of the wide geographic dispersal of the mother clade H7c in both Western Asia and Europe (estimated TMRCA of over 7,000 YBP; Behar et al., 2012), and the documented presence in the Levant of the daughter branch H7c1, which includes the Druze samples and at least one individual of Sephardic origin, a Levantine source for the precursor of H7c2 is a possibility. However, considering that the sister clade H7c3, as well as some of the H7c1 samples, trace their ancestry to Northern Europe, it is difficult to reach a conclusion based on this evidence. If the absolute time estimate for H7c2 is correct, this timing would also support a non-European origin for the maternal ancestress of the local Ashkenazi H7c2 mutation, because it dates to the early period of the Jewish diaspora (200–300 CE; i.e., it pre-dates 650 CE) when the vast majority of Jews were found outside of Europe (see Appendix). However, as noted, absolute time estimates from genetic mutations rates are problematic and cannot presently be relied upon to disambiguate origin. Brotherton and colleagues (2013), for example, using dated haplogroup H genomes to calculate mutation rates, found a mutation rate 45% higher than current estimates for human mitochondria.

H7e

In contrast with H7c2 and H7j, which were found to be exclusively Jewish, H7e included a few individuals of European ancestry with no known Jewish ancestry. H7e was also the largest of the predominately Jewish clusters within H7, with 54 of the 63 individuals of self-described certain Ashkenazi Jewish. Behar and colleagues (2012) dated H7e to the 5^{th}–6^{th} Century CE, but, as with other examples noted, use of a different mutation rate or a high standard deviation means the cluster could either predate or postdate the critical 650 CE time boundary. We did not identify any individuals carrying only one of the defining mutations of H7e (8026T and 9527T), consistent with earlier work by Atzmon et al. (2010). H7 itself has been estimated to be 8890 years old (Behar et al., 2012), many thousands of years older than H7e. Overall, no conclusion can be drawn about the origin of H7e from looking at the haplotypes upstream.

Figure 1


H7c1
16265G,
(16093Y)
11
5001R
1
H7c
6269A
H7c2
13959T
20
709A,
6935T
1
H7c3
9329A.
15459T
3
H7 Ancestral Node
263G, 315.1C,
750G, 1438G,
4769G, 4793G,
8860G,15326G,
16519C
H7e
8026T,
9527T
63
see figure 2
H7 - 1700C > H7j
152C,
573.1C,
1700C
3
573.2C
1
H7j1
11137C
10

Of the 63 individuals with H7e, 31 belonged to the ancestral cluster and carried only the defining mutations of the clade, 8026T and 9527T. In addition, 28 of these 31 individuals were either self-described certain Ashkenazi Jewish or were highly likely to have Ashkenazi roots based on the information provided about their direct maternal lines. For two individuals, there wasn't sufficient information to determine whether they had Ashkenazi roots, and one individual had no known Ashkenazi roots. None of those belonging to the Ashkenazi cluster were aware of Sephardic or other Jewish roots.

Ashkenazi Jewish H7e

In addition to the ancestral cluster in H7e, a number of distinct Ashkenazi clades within H7e were found. The cluster with the greatest internal diversity, which we tentatively labeled H7e1, was identified by the additional mutation 8994A in the coding region. All known members of H7e1 reported Ashkenazi ancestry on their maternal lines. The sequence most distant from the ancestral cluster had three additional mutations (Figure 2). The deep nesting provided evidence of the longevity of H7e among Ashkenazi Jews. An additional large Ashkenazi cluster, tentatively labeled H7e2, was identified by the mutation 12651A.

In total, 84% of the samples belonging to H7e had or highly likely had Ashkenazi Jewish roots on their direct maternal lines. The geographic distribution of these individuals in the ancestral cluster encompassed practically all of the countries in which Ashkenazi Jews lived at the beginning of the 20th Century, from Germany and Austria in Western Europe, through Poland and the Baltic states in Central Europe, to Moldova, Ukraine and Belarus in Eastern Europe. Furthermore, within the Ashkenazi subclades of H7e, distinct regional patterns of distribution were discernable, with disproportionate numbers reporting Lithuanian ancestry (60%) in H7e1 (8994A) and Polish ancestry (50%) in H7e2 (12651A).

The wide distribution of the ancestral cluster along with the more regional distribution of the subclades indicate that H7e entered the Ashkenazi gene pool at a relatively early stage in the history of the haplogroup. The emergence most probably occurred no later than during the 9th and 10th centuries during the formative stages of Ashkenazi Jewry and prior to the movement eastwards to Central and finally to Eastern Europe.

Non-Jewish H7e

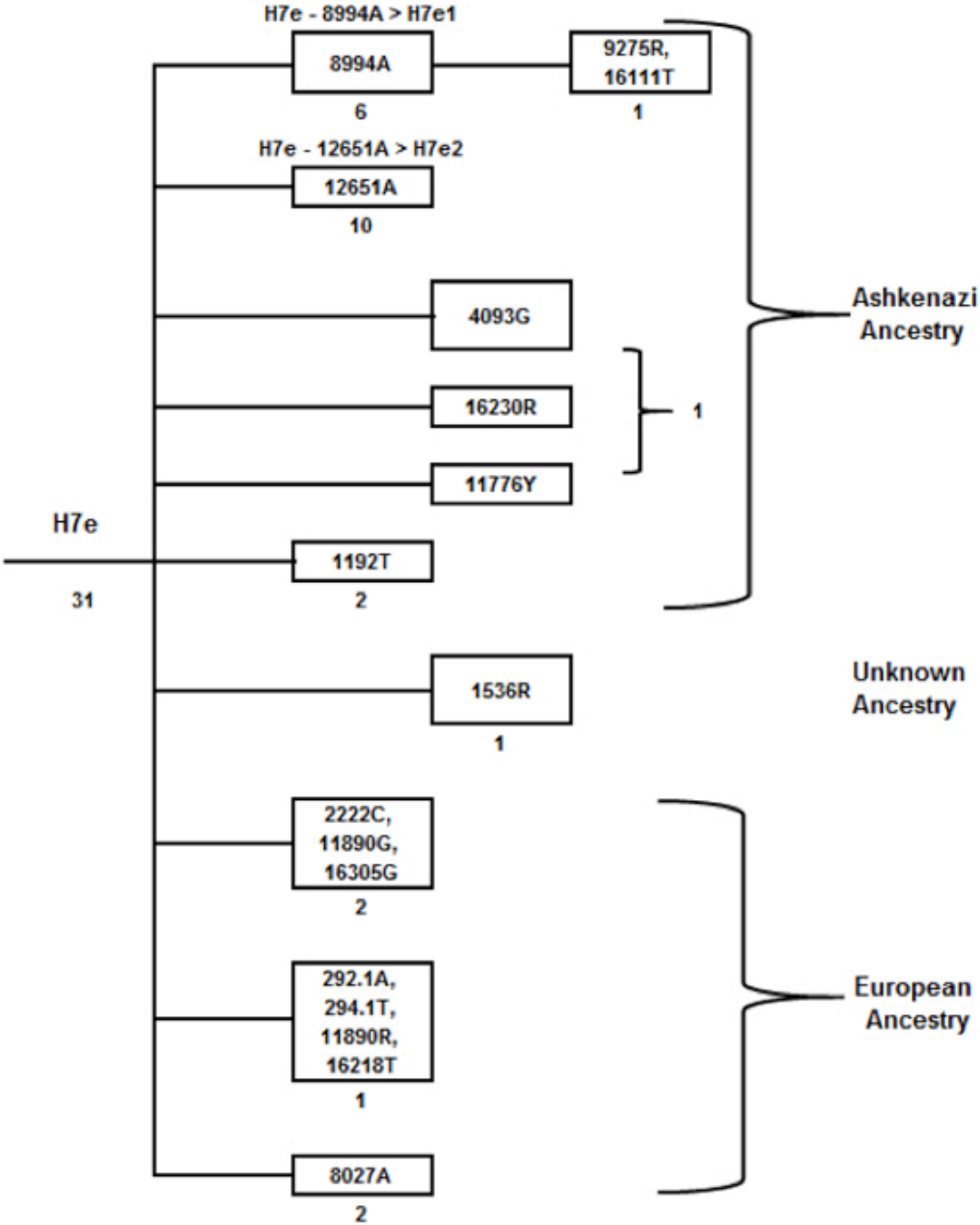
Of the 63 H7e individuals, six had no known Ashkenazi ancestry (~10%), including two who can trace their ancestry back to Germany and one to the island of Susak in Croatia. The remainder could not trace their ancestry beyond colonial America. Another three individuals are unlikely to have Ashkenazi ancestry (~5%).

A striking aspect about the non-Jewish H7e results is that they were found to be a considerable genetic distance from the ancestral cluster and separated by several mutations (see Figure 2). One sequence had four possible independent mutations (16218T, 292.1A, 294.1T, 11890R), and two sequences had three mutations (2222C, 11890G, 16305G). Furthermore, these clusters did not nest within the existing Jewish subclades of H7e, nor did those nearer to the ancestral cluster with no known Jewish roots. There seems to be a clear distinction between those belonging to the subclade with Ashkenazi Jewish roots and those without Ashkenazi Jewish roots, bar one member of the ancestral cluster with no known Jewish roots (< 4% of the ancestral cluster) The non-Jewish samples also show greater genetic diversity than the Jewish samples.

Discussion

The current work identified three clades and several subclades of H7 as predominantly Jewish. One of these (H7j) was previously undiscovered, and the others (H7e, H7c2) had not previously been identified as mainly Jewish. We focused on the European haplogroup H, rarely discussed within Ashkenazi genetics, to gain insight into early European Jewish maternal origins.

Figure 2


H7e - 8994A > H7e1
8994A
6
9275R, 16111T
1
H7e - 12651A > H7e2
12651A
10
4093G
16230R
11776Y
1
Ashkenazi Ancestry
H7e
31
1192T
2
1536R
1
Unknown Ancestry
2222C, 11890G, 16305G
2
292.1A, 294.1T, 11890R, 16218T
1
8027A
2
European Ancestry

The largest group was H7e, with 63 individuals. This reflects the largest collection of complete H7e sequences reported to date; adding to the previous five sequences available on GenBank. At least two regionally distinct subgroups were newly found within H7e. The relatively large sample enabled several patterns to be revealed: 1) The bulk of H7e individuals have Ashkenazi maternal origins. 2) The geographic origins of Ashkenazi H7e encompassed all regions in which Ashkenazim were found including Germany and Austria in Western Europe, Poland and the Baltic states in Central Europe, and Moldova, Ukraine and Belarus in Eastern Europe, with regional subclades apparent. 3) Some H7e sequences were found in individuals who knew of no Jewish ancestry. 4) The Non-Jewish sequences showed rich nesting and several mutational differences from ancestral H7e. And, 5) the non-Jewish clusters showed no overlap with Jewish subclades. Taken together, these findings strongly implicate the introgression of a mitochondrial lineage either from or into the Jewish gene pool that occurred early in the settlement of European Jews. This was followed by no further genetic contact between the two groups. Genetic isolation led to separate expansions, especially among the Ashkenazi as they made their way deep into Eastern Europe.

One challenge facing research into Jewish maternal lineages has been their distinctiveness, which makes their origins difficult to determine. That is, many maternal lineages found among Jewish populations, despite having significant coding region variability, are restricted solely to the Jewish subgroup to which they are found in. In H7e, on the other hand, we found distinct evidence of both Ashkenazi Jewish and European non-Jewish maternal lineages with clear relationships based on coding region variability. Thus we can see genetic evidence of an oft-speculated but rarely seen early exchange, followed by independent development, in the gene pool between Jewish and non-Jewish groups.

But in which direction was the early genetic contribution? The dominance of haplogroup H as an early European rather than Near Eastern haplogroup may favor the hypothesis that one woman belonging to Haplogroup H7e converted to Judaism and married into the Jewish community. The predominance of Jewish individuals within the ancestral cluster would, in this view, be explained by the Ashkenazi bottleneck and subsequent population boom (Carmi et al., 2014) which resulted in an inflated number of Ashkenazi Jewish women carrying the ancestral version of H7e than in the general European population.

One is also tempted to speculate that the non-Jewish European origin of H7e was German. This possibility is consistent with the fact that, of the few individuals without Jewish roots, two could trace their distant ancestry back to Germany. In addition, Ashkenazi Jewish history considers settlement in Germany to have occurred before expansion to Eastern European regions. If this is the case, then H7 is younger than previously thought, because there is practically no evidence of a Jewish presence during the 7th and 8th centuries in the Rhineland area (see Appendix).

A second possibility consistent with an older age for H7e is a European origin in Italy or Southern France. The Jewish presence in the Rhineland area, and later in central Europe, is considered the outcome of the migration of Jews from Southern Europe that began in the 9th and 10th centuries (Botticini & Eckstein, 2012). The gene flow, however, could have occurred in either direction: for example, non-Jewish French women marrying newly arriving Near Eastern Jewish men or Jewish women arriving to Italy from the Near East and leaving the Jewish community. Origin of H7e in Italy or Southern France would require an explanation for why all traces of the haplogroup have vanished from those areas. Such an explanation may not be hard to find. In general, many — perhaps most — haplogroups have likely vanished from existence; the unusual situation of the Ashkenazi extreme bottleneck and subsequent population explosion allowed otherwise extinguished haplogroups to survive in select demographics.

Finally, despite the predominance of haplogroup H in Europe and the other factors suggesting a European origin, we cannot definitively rule out the other extreme: that the ancestress of H7e was herself part of the Jewish community in antiqui-

ty. Regardless of where geographically the women were when the mutations of H7e arose, they still could have arisen in women whose ancestors were Jewish before leaving the Near East. H7 and other H clades could nonetheless have been in the Near East at the right times even if they predominately expanded in Europe. In this view, the small number of non-Jewish individuals belonging to H7e represents the descendants of women who left the Jewish community relatively early on in the history of the subclade. This would include the German, Croatian, and Colonial American participants in our study.

The present work also uncovered a small new clade tentatively labelled H7j and identified the previously known H7c2 group as Ashkenazi Jewish. Neither had any non-Jewish affiliation. The small sizes of the clusters may have precluded any minor non-Jewish presence from being detected, the small clusters may have vanished in all but the large Ashkenazi population, or the mutations characterizing these branches may simply have arisen among the isolated Ashkenazi communities while in Europe. We favor the latter hypothesis. Regardless, it is important to note that an ancient Near Eastern source for the precursors of H7c2 or H7j is possible under any of the hypotheses. We also found interesting patterns in the smaller H7j and H7c2 clusters. One cluster contained several, non-overlapping, shallow branches that emerged contemporaneously, reflecting a relatively new clade in a large population. The other pattern revealed a possible in-progress vanishing of the ancestral group, which may soon be lost to history and lead to missing links in the phylogenetic tree.

As analysis of H7 clades illustrates, determining the direction of gene flow with any degree of certainty is difficult, even when sequences belonging to non-Jewish populations are found (as for H7e). The problem is even greater when a mitochondrial lineage is restricted exclusively to Ashkenazi Jews, as often occurs. Consequently, it is notable that Costa and colleagues (2013) nonetheless concluded that 80% of Ashkenazi maternal ancestry is due to the assimilation of mtDNAs indigenous to Europe, most likely through conversion. We feel this conclusion is premature and goes beyond the available evidence for several reasons: the intricacies of Jewish history are often overlooked, the methodology of looking at the immediate ancestral nodes is not always conclusive, time estimates that can be grossly inaccurate are often relied on too heavily, and confusion exists between where an individual lived when a de novo mutation arose and that person's origins. We provide an example and brief elaboration from the Costa et al., 2013 paper to illustrate. We belabor the point because of the importance of concluding such a definitive maternal origin for the vast number of Ashkenazi haplogroups.

The haplogroups surveyed by Costa and colleagues (2013) may have arisen in Europe between the last glacial period and the Neolithic as maintained. However, when, considering the complex history of migration within the Mediterranean basin over the last 3,000 years, as well as Jewish history (see Appendix), it is apparent that where a haplogroup first arose many thousands of years earlier need not have any bearing on where and when a specific distinctive mitochondrial haplogroup first emerged among Jewish populations. Furthermore, a sizeable portion of the Mediterranean–Hellenistic Jewry of antiquity was comprised of converts to Judaism rather than descendants of the Iron-Age Israelites. While the majority of these converted in the land of Israel prior to 65 CE, they undoubtedly included some descendants of merchants, colonists, and troops with roots tracing back to Mediterranean Europe, which could explain some of the European admixture found amongst the Jewish populations descending from the Mediterranean-Hellenistic Jewry of antiquity based in the Eastern Mediterranean.

For a specific example, consider the often discussed haplogroups K1a1b1a and K1a1b1a1 among Ashkenazi Jews. Costa and colleagues (2013) used maximum likelihood to estimate that K1a1b1a dates to approximately 4,400 YBP and K1a1b1a1 to 2,300 YBP. To place these results in their historical perspective, 2,300 YBP predates the dispersal of the Jewish population from the Levant to Europe, and 4,400 YBP predates by more than 1,000 years the earliest documented mention of the name "Israel" in historical record (the Mer-

neptah Stele, dated to 1209 BC). As they estimate the parent clade K1a1b1 to be over 10K years old, in the interim ~6,000 years between the appearance of K1a1b1 and the appearance of K1a1b1a, the maternal lineage could have migrated to and from the Levant on numerous occasions (in a manner similar to the movement pattern of H7c1). As noted earlier, prior to the Arab conquest in the 7th century CE the Western and Eastern sides of the Mediterranean basin were as well, if not better, connected to each other than the Western Mediterranean was to parts of Northwestern Europe. When considering the age of the haplogroup, its presence (however limited) among Sephardic Jews and its apparent absence in non-Jewish populations (Costa et al., 2013; Behar et al., 2006) all seem to indicate that a Levantine origin is far more likely for K1a1b1a than a European one, regardless of where K1a1b1 first originated.

Turning attention to mtDNA mutation rates, our finding of early exchange between the European and Jewish gene pools in haplogroup H mtDNA (H7e) suggests that the rates of mutations are much faster than commonly assumed. They are closer to those estimated using pedigrees. Madrigal and colleagues (2012) calculated a mutation rate of 1.24×10^{-6} per site per year in an analysis of individual family pedigrees from a well-documented population in Costa Rica, a rate three times faster than those commonly derived from phylogenies. The distinctiveness of Ashkenazi Jewish maternal lineages and their isolation from non-Jewish maternal lineages, coupled with a rapid population explosion and the relatively well-documented history of Ashkenazi Jewry, may provide a further basis for grounding the widely varying mutation rates offered by different sources.

Finally, we can reconsider the high degree of European admixture (30%–60%) observed among Ashkenazi, Sephardic, Italian, and Syrian Jews (Atzmon et al., 2010) in autosomal DNA studies, as well as the higher proportion of European admixture among North African Jews compared with non-Jewish North African populations (Campbell et al., 2012). Part of this clearly reflects limited more recent European admixture, hence the elevated levels of European admixture when comparing Ashkenazi to Sephardic Jews or Moroccan to Djerban Jews. However, part undoubtedly reflects the legacy of the Mediterranean and the movement of peoples around the Mediterranean basin long before Christian Southern Europe become isolated from the Islamic Levant and North Africa, and results from conversions to Judaism prior to 65 CE in the Hellenistic and then Roman Levant and North Africa.

Little is known about the earliest days of settlement of the Ashkenazi Jews in Europe. Research into Jewish population genetics holds the promise of illuminating migrations and expansions that are poorly understood due to the scarcity of reliable historical sources. We believe we have provided one of the clearest views of this early period through a branch of maternally inherited mitochondrial DNA haplogroup H that strongly implicates gene flow between the Ashkenazi and non-Jewish European populations pre-dating the Ashkenazi expansion throughout Central and Eastern Europe. We focused on the most prevalent haplogroup in Europe, which also contains subclades found almost exclusively among Ashkenazi Jews, to provide further insight into the origins of the European Jewish communities. We found gene flow within haplogroup H7, evidence that will be beneficial in assessing the origin of other mitochondrial subclades found among Jewish groups.

Acknowledgments

The authors gratefully acknowledge the constructive critiques provided by Leah Larkin, Ian Logan, and two anonymous reviewers as well as the assistance of Jacques Beaugrand, administrator of the H7 MtGenome Project.

Appendix A. A Brief Consideration of Mediterranean and Jewish History

Historical considerations in mtDNA genetic studies tend to focus on prehistoric Europe because of the ages of many haplogroups and, in particular, the last glacial maximum and its impact on human migrations (Roostalu et al., 2007). Often overlooked, , however, is that following these events many thousands of years ago, human migration continued unabated and, with it, the corresponding gene flow between different parts of Europe, Western Asia, and North Africa (e.g., Brotherton et al., 2013 re Haplogroup H in Europe).

One of the most important facilitators of migration between these geographical areas was the Mediterranean. As Abulafia (2003) pointed out, thanks to the ease of movement across the open sea, lands far removed from each other enjoyed vibrant trading, cultural, and political ties. Furthermore, from the Mediterranean, access could be gained to the European network of big rivers, such as the Danube and the Rhine, further facilitating the movement of goods and people from the Mediterranean basin inland into Central Europe. There is no doubt that this movement around the Mediterranean basin has very ancient roots. Archaeological sites in Israel reveal a Stone Age culture quite similar to that known in the Western Mediterranean from the limestone caves of Spain, France, and Northern Italy (Suano, 2003).

The Mycenaeans in the 14th century BCE were the first to start intensively traversing the Mediterranean carrying trade between the Aegean and the Levantine coastal cities, thus linking these regions to the central Mediterranean and, on occasion, Iberia. Permanent settlements of Mycenaeans have been identified on the coast of southern Italy, in Sicily, and in Sardinia (Torelli, 2003). The commercial traffic of the Mediterranean throughout the pre-Roman age was marked by colonial settlement as much as by mercantile contact. Following the collapse of the Mycenaean empire and the rise of classical Greece and Phoenicia, the trade rivalry between the Greeks and Phoenicians and the ensuing battle over the Mediterranean trading routes between 1,000 BCE and 300 BCE led to the development of a wide ranging network of trading settlements and colonies. Colonies in Carthage and the ring of emporia in Libya, Motya, and Soluntum in Sicily; the harbors in Sardinia; and the bases and trading stations at Ibiza in the Baleric Islands, Cadiz beyond the straits of Gibraltar, and along the Moroccan Coast allowed the Phoenicians to dominate many of the trade routes straddling North Africa, Iberia, and the Levant. The Greeks as well as the Etruscans developed rival trading routes covering much of Southern Europe, the Adriatic, the Black Sea, and Asia Minor (Torelli, 2003).

The key period of Mediterranean unification occurred, however, under the rule of imperial Rome. For a period of roughly 800 years (300 BCE–500 CE) the whole Mediterranean was politically unified. As Rickman (2003) stated, "it is hardly surprising that a sea which the Romans, and the polyglot populations under their control had so thoroughly made their own …. should witness not just the circulation of goods, but also of people". Military conquests during the Republic (300–100 BCE) and the expansion of the Roman Empire brought to the Italian peninsula significant economic migration of free immigrants as well as slaves from Gaul, Hispania, Germania, Magna Graecia, Asia Minor, Phoenicia, Egypt, and North Africa (Noy, 2000; Scheidel, 2004). Scheidel (2004) estimates that around 2 million people immigrated to Rome just during the last two centuries BCE while, according to Noy (2000), over 10% of foreigners buried at Rome came there from North Africa, and most were civilians rather than associated with the military (see Killgrove, 2010, 2013). The movements of people were not just to Rome. The names of the units stationed on Hadrian's Wall reveal how widely Rome recruited its auxiliary regiments, from Spain, Gaul, Germany, the lands along the Danube, Asia Minor, Syria, and North Africa (Vindolanda, 2016).

Jewish history is intertwined with Mediterranean history. The formative stages of the Jewish diaspora occur during the period of the Mare Nostrum (or 'our [Roman] sea'). There is a tendency to confuse the Iron-Age Israelites of the 8th and

9th centuries BCE with the Jewish population living in the Roman province of Judea nearly 1,000 years later just prior to the great revolt of 65–70 CE, however, while undoubtedly some of those living in Judea as Jews during the 1st century CE were the genetic descendants of the inhabitants of the ancient kingdoms of Israel and Judah, many others were not. The four centuries following the Babylonian conquest of Judah in 586 BCE had seen major political and demographic changes taking place in the land of Israel. Faust (2012) has persuasively shown that, based on the archaeological evidence, Judah experienced drastic demographic decline due to the war, subsequent famine, and epidemics that followed the conquest. Continuity in the following centuries with the Iron Age society of Judah was limited. There were survivors, and some of the population exiled to Babylon must have returned, but population recovery in the region must have also been triggered by new settlers from neighboring regions (Faust, 2012). Following its conquest by Alexander the Great in 332 BCE, Judea was no longer merely a buffer state between Egypt and Mesopotamia; it now formed the eastern edge of what was quickly becoming a pan-Mediterranean empire — the Roman 'Mare Nostrum'. By 63 BCE, Judea was a client state of Rome and by 6 CE a Roman province.

In Goodman's (1994) thorough research into proselytes and proselytizing to Judaism during the period of the Roman Empire, he concluded that there is evidence that prior to 65 CE, converts made up a significant proportion of the Jewish population and that Jews accepted as proselytes those gentiles who applied to join their number, although they did not feel compelled to encourage such conversions. As examples, Goodman (1994) referred to the spread of Jewish settlement in the diaspora, the increase in the population of Judea apparent from archaeological survey, and Josephus' recording of the conversion en masse of neighboring populations such as the Idumeans and the Ituraeans by the Hasmonaean dynasty.

In the post-70 CE period, ambivalence by Rabbinical authorities towards the proselytization of gentiles meant that conversion to Judaism was far less common, although there is some evidence of proselytes to Judaism all the way through into the medieval period (Goodman, 1994). This was especially true after the failed Bar Kokhba rebellion during Hadrian's rule and the passage of legislation by Hadrian and his successors against the circumcision of non-Jews, the special Jewish tax (the fiscus Judaicus), and a series of Roman laws in the 4th and 5th centuries prohibiting conversion to Judaism, particularly by Christians. Furthermore, as Goodman (1994) pointed out, some conversions to Judaism probably took place to facilitate marriage. Considering the patriarchal nature of both Jewish and Roman societies, as well as the prohibition on circumcision that prevented men (but not women) from converting, many of the converts to Judaism to facilitate marriage were likely women.

How many of these conversions would have taken place in Europe? As can be seen in Table 1 based on the estimates of Botticini and Eckstein (2012), prior to 65 CE the majority of the Jewish population throughout the Middle East and the Mediterranean basin were located in the lands of Israel, Mesopotamia, Persia, and North Africa (mainly Egypt), while the number of Jews in Western Europe was relatively small and by 650 CE was negligible (~1,000). Thus the vast majority of conversion to Judaism during this period must have occurred outside of Europe in the Levant, Egypt, and Mesopotamia.

Furthermore, in a detailed study by Toch (2005) of Jews in Europe between 500–1050 CE, he concluded that between the mid-7th and mid-8th centuries, no source mentions Jews in Frankish lands (now France and Germany). Only in the 8th and 9th centuries was there evidence of growing numbers of Jews in the South of France, while in the 9th and early 10th centuries, brief hints attest to itinerant merchants in Germany. Toch (2005), therefore, concluded that no continuity could be assumed between the Jews of the Roman Empire and the Ashkenazi Jewish communities of the Middle Ages.

From a genetic perspective, based on this historical overview, maternal lineages restricted to Jew-

ish populations that pre-date 650 CE are highly unlikely to have originated in either Western or Eastern Europe, given the miniscule numbers of Jews in these regions during this period.

Table 1. Jewish population estimates in 65 CE and 650 CE (as per Botticini and Eckstein, 2012).

Region	c. 65 CE	c. 650 CE
Land of Israel	2,500,000	100,000
Mesopotamia and Persia (including the Arabian Peninsula)	1,000,000	700,000–900,000
North Africa (mainly Egypt)	1,000,000	4,000
Syria and Lebanon	200,000–400,000	5,000
Asia Minor and the Balkans	200,000–400,000	40,000
Western Europe (including Italy, France Germany, and Iberia)	100,000–200,000	1,000
Eastern Europe	–	–

References

Abulafia D (2003) What is the Mediterranean? In The Mediterranean in History, D Abulafia, ed. Getty Publications, Los Angeles, CA. pp. 11–32.

Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E, Ostrer H (2010) Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern ancestry. American Journal of Human Genetics 86: 850–859.

Bedford FL (2012) Sephardic signature in haplogroup T mitochondrial DNA. European Journal of Human Genetics 20: 441–448.

Bedford FL, Yacobi D, Felix G, Garza FM (2013) Clarifying mitochondrial DNA subclades of T2e from Mideast to Mexico. Journal of Phylogenetics and Evolutionary Biology 1: 121. doi:10.4172/2329-9002.1000121

Bedford FL, Yacobi D (2015) On two Jewish clades in mitochondrial DNA. European Journal of Human Genetics 23: 993-994. doi: 10.1038/ejhg.2014.231.

Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, Tzur S, Pereira L, Amorim A, Quintana-Murci L, Majamaa K, Herrnstadt C, Howell N, Balanovsky O, Kutuev

I, Pshenichnov A, Gurwitz D, Bonne-Tamir B, Torroni A, Villems R, Skorecki K (2006) The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. American Journal of Human Genetics 78: 487–497.

Behar DM, Yunusbayev B, Metspalu M (2010) The genome-wide structure of the Jewish people. Nature 466: 238–242.

Behar DM, Van Oven M, Rosset S (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. American Journal of Human Genetics 90: 675–684.

Botticini M, Eckstein Z. (2012) The Chosen Few: How Education Shaped Jewish History, 70–1492. Princeton University Press, Princeton, NJ.

Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Adler CJ, Richards SM, Der Sarkissian C, Ganslmeier R, Friederich S, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1764.

Campbell CL, Palamara PF, Dubrovsky M, Botigué LR, Fellous M, Atzmon G, Oddoux C, Pearlman A, Hao L, Henn BM, Burns E, Bustamante CD, Comas D, Friedman E, Pe'er I, Ostrer H (2012) North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. Proceedings of the National Academy of Sciences USA 109: 13865–13870.

Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, Guha S, Upadhyay K, Ben-Avraham D, Mukherjee S, Bowen BM,

Thomas T, Vijai J, Cruts M, Froyen G, Lambrechts D, Plaisance S, Van Broeckhoven C, Van Damme P, Van Marck H, Barzilai N, Darvasi A, Offit K, Bressman S, Ozelius LJ, Peter I, Cho JH, Ostrer H, Atzmon G, Clark LN, Lencz T, Pe'er I (2014) Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nature Communications 5: 4835. doi:0.1038/ncomms5835.

Costa MD, Pereira JB, Pala M, Fernandes V, Olivieri A, Achilli A, Perego UA, Rychkov S, Naumova O, Hatina J, Woodward SR (2013) A substantial prehistoric European ancestry among Ashkenazi maternal lineages. Nature Communications 4: 2543 doi:10.1038/ncomms3543.

Faust A (2012) Judah in the Neo-Babylonian Period: The Archaeology of Desolation. Society of Biblical Literature, Atlanta, GA, USA.

Goodman M (1994) Mission and Conversion: Proselytizing in the Religious History of the Roman Empire. Oxford University Press, Oxford, England.

Killgrove K (2013) Biohistory of the Roman Republic: the potential of isotope analysis of human skeletal remains. Post-Classical Archaeologies 3: 41–62.

Killgrove K (2010) Identifying immigrants to Imperial Rome using strontium isotope analysis. In Roman Diasporas: Archaeological Approaches to Mobility and Diversity in the Roman Empire, ed. H Eckardt. Journal of Roman Archaeology supplement 78, ch. 9, pp. 157–174.

Madrigal L, Castri L, Melendez-Obando M, Villegas-Palma R, Barrantes R, Raventos H, Pereira R, Luiselli D, Pettener D, Barbujani G (2012) High mitochondrial mutation rates estimated from deep-rooting Costa Rican pedigrees. American Journal of Physical Anthropology 148: 327–333.

Noy D (2000) Foreigners at Rome: Citizens and Strangers. Classical Press of Wales, Swansea, Wales.

Ostrer H, Skorecki K (2013) The population genetics of the Jewish people. Human Genetics 132: 119–127.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, Tamm E, Karmin M, Reisberg T, Hooshiar Kashani B, Perego U A, Carossa V, Gandini F, Pereira JB, Soares P, Angerhofer N, Rychkov S, Al-Zahery N, Carelli V, Sanati MH, Houshmand M, Hatina J, Macaulay V, Pereira L, Woodward SR, Davies W, Gamble C, Baird D, Semino O, Villems R, Torroni A, Richards MB (2012) Mitochondrial DNA signals of late glacial recolonization of Europe from Near Eastern refugia. American Journal of Human Genetics 90: 915–924.

Pike DA, Barton TJ, Bauer SL, Kipp EB (2010) mtDNA haplogroup T phylogeny based on full mitochondrial sequences. Journal of Genetic Genealogy 6: 1–24.

Pike DA (2006) Phylogenetic networks for the human mtDNA haplogroup. Journal of Genetic Genealogy 2: 1–11.

Rickman G (2003) The Creation of Mare Nostrum: 300 BC–500 AD. In The Mediterranean in History, D Abulafia, ed. Getty Publications, Los Angeles, CA. pp. 127–154.

Roostalu U, Kutuev I, Loogvali EL, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, Villems R (2007) Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Molecular Biology and Evolution 24: 436–448.

Scheidel W (2004) Human mobility in Roman Italy, I: the free population. Journal of Roman Studies 94: 1–26.

Shlush L., Behar D, Yudkovsky G, Templeton A, Hadid Y, Basis F, Hammer M, Itzkovitz S, Skorecki K (2008) The Druze: a population genetic refugium of the Near East. PLoS ONE 3: e2105. doi:10.1371/journal.pone.0002105.

Suano M (2003) The first trading empires: prehistory to c. 1000 BC. In The Mediterranean in History, D Abulafia, ed. Getty Publications, Los Angeles, CA. pp. 67–98.

Toch M (2005) The Jews in Europe 500–1050. In The New Cambridge Medieval History, vol. I, 500–1050, P Fouracre, ed. Cambridge University Press, Cambridge, England. pp. 547–570.

Torelli M (2003) The battle for the sea routes: 1000–300 BC. In The Mediterranean in History, D Abulafia, ed. Getty Publications, Los Angeles, CA. pp. 99–126.

Van Oven M (2015) PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Science International Genetics 5: e392–e394.

Vindolanda Tablets Online database, http://vindolanda.csad.ox.ac.uk/, accessed 26 Jul 2016.

Journal of Genetic Genealogy 8(1):35-37, 2016

Editorial

The History of Genetic Genealogy and Unknown Parentage Research: An Insider's View

CeCe Moore
http://www.TheDNADetectives.com

When the last issue of JOGG was published in the Fall of 2011, using genetic genealogy to identify recent unknown parentage was in its infancy. Genetic genealogists were squarely focused on using DNA to learn more about our distant ancestors and conquer our genealogical brick walls. Around that time, I became aware of a whole category of people who were denied knowledge of their genetic origins and the joy of building a family tree, at least one tied to their biological ancestors: those with unknown parentage. There were also a surprising number of genealogists taking DNA tests and discovering, unexpectedly, that half of their trees, often the results of decades of research, was not their true genetic pedigree.

As a genealogist, I feel strongly that everyone has the right to explore their genetic origins, research their ancestors, and participate in the popular hobby of genealogy. As I learned, for many adoptees and others of unknown parentage, this had proven impossible. It seemed obvious that genetic genealogy could help them.

Prior to the introduction of commercial autosomal DNA testing for genealogy in late 2009, to my knowledge, there were only a couple of men who had resolved their unknown paternity using Y-DNA testing. Notably, Richard Hill was profiled in the *Wall Street Journal* for his discovery of his biological paternal heritage in 2009 (Naik, 2009). He later self-published the book *Finding Family: My Search for Roots and the Secrets in My DNA* (Hill, 2012). But for the majority, resolving unknown parentage was an elusive goal. In 2011, the 23andMe and FTDNA autosomal DNA (atDNA) databases were extremely small, and AncestryDNA's had not even been created. This presented a challenge. How could we best use these data to help those who had no other information on their roots?

A group of adoptees, traditional search angels, and a genetic genealogist (myself), banded together to devise methods to harness the vast amount of genetic data generated by atDNA testing (even with the smaller databases of the time) for unknown parentage searches. In those early days, the matches were almost always distant, 4th to 6th cousins and beyond, so in most cases, segment triangulation made sense.

Segment triangulation entails grouping matches together who all share atDNA with one another on the same or overlapping segment and looking for a common ancestral line among them. All of those matching on the same/overlapping segment are reasoned to share an ancestral line and, so then, should the person of unknown parentage. When common ancestors were successfully identified, teams of volunteers and adoptees spent hundreds or often thousands of hours combining the ancestral lineages of the matches and building huge family trees backward and forward in time, hoping to trace to the present to find a person who was in the right place at the right time to be the unknown parent. Citizen scientists created tools specifically to help the search community make sense of the vast amounts of data at our fingertips. Those tools ultimately benefited the entire genetic genealogy community. However, the work was grueling, and we saw few success stories. The more recent discoveries of "pile-up regions" and the true depth of atDNA matching explain some of the difficulties we unknowingly faced at the time.

It quickly became apparent to us that predicted second cousins were the "sweet spot" for identifying birthparents. If two people share about 3% (roughly 212 cM) of their atDNA, then there is a good chance that they share a set of great grandparents. Tracing the descendants of the eight great grandparents forward in time, unsurprisingly, leads to potential birth parents. As the

databases grew, we quickly saw more cases being solved in this way and with much less emphasis on segment triangulation and building huge, speculative family trees.

In mid-2012, Ancestry.com launched their atDNA service, which had a significant effect on the way unknown parentage searches were resolved. What we had needed all along was more pedigree data for those sharing DNA. Matching segment data without the family trees of the matches was virtually useless, so the founding members of the search community had spent much of our time tracking down or building those trees. Since Ancestry.com had long been in the business of collecting family tree data, they had a unique opportunity to correlate the pedigrees directly with the atDNA data they were quickly accumulating. In many cases, it was no longer even necessary to contact the match, making the work considerably faster and simpler. Triangulating family tree data, rather than segment data, was much more attainable and very successful in identifying shared ancestors.

Later, AncestryDNA's "Shared Ancestor Hints" automated identification of common ancestors, giving us a new, extremely valuable tool in our pursuit. Instead of spending many hours manually searching for common ancestors among the subject's matches, "mirror trees" could be created and attached to the DNA results of a person of unknown parentage to automatically search for common ancestors in the family trees of their DNA matches.

What is a mirror tree? A mirror tree is built based on the pedigree of a DNA match to the person searching. Recreating it, or even better, being invited to editor status by the owner, allows the searcher to attach their DNA results, as if the trees was their own. When it works, the Shared Ancestor Hints can quickly identify which branch of the match's tree is in common with the subject of unknown parentage by finding third parties who share both DNA with the adoptee and ancestors with the mirrored match. Speculative trees are also a very useful for unknown parentage work. By building out the family tree of a candidate birth parent as deeply as possible on all ancestral lines, one can usually determine whether the DNA of the searcher and the tree of the prospective birth parent correlate well via the Shared Ancestor Hints.

In the early days of adoptee searching, we used to refer to it as "being struck by lightning" if an adoptee received a close family match in the databases. However, with the incredible growth of these databases over the last couple of years, this has become more and more common. In fact, today, in the DNA Detectives Facebook group alone (https://www.facebook.com/groups/DNADetectives/), we see such matches every day. When large batches of new AncestryDNA matches load, we will often see multiple half-sibling, aunt/uncle, and first-cousin matches among the members of the group. Supporting my anecdotal experience, a recent survey of people who DNA tested to find birth family found that 90% were matched to a 3rd cousin or closer immediately upon receiving their results (Bettinger, 2016). Thus, I believe that our genetic genealogy databases have hit critical mass, at least for those with deep roots in the United States, and even for those whose great grandparents were not all immigrants. It is difficult to fathom how this could be true since the total number of testers is only roughly 1% of the U.S. population, but what we are witnessing with our own eyes cannot be denied.

Experienced genetic genealogists have often joked that some people are under the misconception that they will take a DNA test and their family tree will automatically generate. In the not too distant future, this is very likely to become a reality, at least to a moderate extent. The databases are currently growing at breakneck speed and, before long, successful birth-family searches and immediate family reunions made possible through DNA testing will be as commonplace as taking a DNA test and matching to a second or third cousin is now. The experts in unknown parentage work will have to find another area on which to focus because, thankfully, the answers long-sought by those of unknown parentage will be easy to come by and these mysteries will take little special skill to unravel.

The development of genetic genealogy methods for unknown parentage searches has been an important and productive effort for our community. This work and the tools created to support it have benefited those searching for immediate biological family as well the genetic genealogy communi-

ty as a whole. Further, the media coverage of these types of cases has significantly increased public interest in our industry, attracted multitudes of new testers, and inspired new genetic genealogists. Undoubtedly, the process has reinforced the concept that learning about one's family history is a valuable and worthy endeavor for all.

About the Author

CeCe Moore is an independent genetic genealogist, an innovator in the field of unknown parentage searches, and an educator in the genetic genealogy community. She is the genetic genealogy consultant and scriptwriter for *Finding Your Roots*, the co-founder of the *Institute for Genetic Genealogy*, and the founder of *The DNA Detectives*.

Conflicts of Interest

CeCe Moore has consulted on a volunteer basis for 23andMe, Family Tree DNA, and AncestryDNA. She declares no conflicts of interest.

References

Bettinger, Blaine. 2016. Preliminary Results of "Adoptee Testing 2016" Survey. https://www.facebook.com/groups/DNADetectives/permalink/1200136676724114/?match=YmV0dGluZ2VyLGJsYWluZQ%3D%3D&__mref=message_bubble.

Hill, Richard. 2012. Finding Family: My Search for Roots and the Secrets in My DNA. Self-published through CreateSpace Independent Publishing Platform, https://www.createspace.com/.

Naik, Gautam. 2009. Family Secrets: An Adopted Man's 26-Year Quest for His Father. http://www.wsj.com/articles/SB124121920060978695

Editorial

The Shared cM Project: A Demonstration of the Power of Citizen Science

By: Blaine T. Bettinger, Ph.D., J.D.
http://www.thegeneticgenealogist.com

Abstract

The Shared cM Project (goo.gl/2uouqz) is a collaborative citizen scientist project created to analyze the ranges of shared centimorgans associated with known genealogical relationships. Between March 2015 and May 2016, members of the genealogical community submitted total shared cM data for almost 10,000 known relationships ranging from parent/child to eighth cousins. The data for each relationship was analyzed to remove extreme outliers, and after determining the minimum reported value, the maximum reported value, and the average for each relationship, a histogram was generated to reveal the distribution between the minimum and maximum reported values. Although susceptible to data entry errors, misattributed parentage, endogamy, pedigree collapse, and company thresholds, these known issues are minimized by the volume of reported values for the majority of relationships. For the first time, genealogists now have observed, non-simulated ranges and distributions of total shared cM data for a wide variety of relationships based on thousands of data points.

Introduction

One of the most common tasks of a DNA test-taker is to derive possible relationships based on the total amount of DNA shared between two genetic matches. Although the three major DNA testing companies (23andMe, AncestryDNA, and Family Tree DNA) each provide a relationship estimate, these estimates can vary and may be based on unclear thresholds. Additionally, relationship estimates may not be available when analyzing shared DNA using a third-party tool.

One of the resources used to predict relationships based on total shared DNA is the Autosomal DNA Statistics page of the International Society of Genetic Genealogy (ISOGG) Wiki (http://www.isogg.org/wiki/Autosomal_DNA_statistics). The page has a variety of sources for relationship predictions, including the table entitled "Average autosomal DNA shared by pairs of relatives, in percentages and centiMorgans," which provides the amount of DNA expected to be shared by individuals having a known genealogical relationship. Although the table assumes exactly 50% inheritance at each generation and thus does not provide an average, it is a very good source for relationship prediction.

However, the ISOGG table does not account for the ranges seen in total shared cM for genealogical relationships. For example, although the expected amount of DNA shared by second cousins is 212.50 cM based solely on 50% inheritance at each generation, the actual average and range for tested second cousins is not provided in the chart. If tested second cousins share 175 cM, is that unusual or is it common? Does that result support a second cousin relationship, or does it suggest another relationship?

There are other sources of data for total shared DNA for genealogical relationships, but these sources are either based in whole or in part on simulated data, or they are created using unknown methodologies and must therefore be used with caution. For example, the "AncestryDNA Matching White Paper" by Ball et al. (31 March 2016; http://dna.ancestry.com/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf) includes Figure 5.2 with the distribution of total shared DNA for a va-

riety of simulated pedigree relationships. Although informative, this data is based on simulated data rather than empirical data. Similarly, the "Average percent DNA shared between relatives" table published by 23andMe contains data based entirely on simulations (https://customercare.23andme.com/hc/en-us/articles/202907170-Average-percent-DNA-shared-between-different-types-of-cousins).

Accordingly, there was a need for empirical data for total shared DNA for genealogical relationships. To fill this need, the Shared cM Project was launched on March 4, 2015. A first analysis of the results was published on May 25, 2016. Discussed herein is an update to the Shared cM Project. This update includes thousands of additional data points, as well as total shared cM data for relationships tested by AncestryDNA. Since AncestryDNA first provided total shared cM data in November of 2015, this update is the first to include this new data.

The data collected by the Shared cM Project is susceptible to several known issues:

- **Data Entry Errors** - Some of the information entered by contributors will include errors resulting from transcribing the data from the testing company or third-party tool and entering the data into the field. For example, for some of the data entries, the longest segment was greater than the total shared cM. Although this was most likely a simple inversion, these data entry errors were completely removed whenever they could be identified. Not all errors, of course, could be reliably identified.

- **Incorrect Relationships** – Some relationships were most likely entered incorrectly, which might be due to misunderstandings of complex genealogical relationships. Other relationship errors are most likely due to misattributed parentage events resulting in the believed relationship being incorrect. For example, with the unedited Aunt/Uncle/Niece/Nephew data, there was a significant cluster around approximately 850 cM, which is indicative of a half-Aunt/Uncle/Niece/Nephew relationship. In other words, there are many unknown half relationships in the data.

- **Endogamy and Pedigree Collapse** – Some relationships will be affected by endogamy and/or pedigree collapse, which will increase the amount of DNA shared by test-takers having a certain genealogical relationship. Although the collection form requests information about known endogamy and pedigree collapse, many contributors will not be aware of the endogamy and pedigree collapse in their tree.

- **Company Thresholds** – Each of the DNA testing companies applies a different matching threshold to maximize the identification of genetic cousins while minimizing false positives. These thresholds may impact the total amount of DNA shared by two test-takers, especially at more distant relationships.

Despite these issues, the volume of hundreds of matches (and, hopefully, thousands of matches in the future) for most relationships in the Shared cM Project are predicted to minimize the impact of these issues on the averages and distributions. Accordingly, the Shared cM Project remains the largest collection of empirical data for total shared DNA for genealogical relationships, and is an example of the power of citizen science.

Methods and Data

Data Collection

Data was collected from participants using Google Forms, which collected the submissions into a spreadsheet. The Google Form (available at goo.gl/qL5BDr) contained data entry fields for required information ("Known Relationship," "Total Shared cM," "Number of Shared Segments," "Endogamy or Known Cousin Marriage" (YES/NO) and "Source" (AncestryDNA, Family Tree DNA, 23andMe, GEDmatch, or Other)), and optional data entry fields ("Longest Block," "Notes," and "Email Address").

A total of 9,891 submissions were made to the Shared cM Project as of May 7, 2016 (beginning March 4, 2015). For analysis, the submissions were downloaded as an Excel spreadsheet.

Initial Data Curation

Because "Known Relationship" was a text entry field, submissions varied considerably regarding the naming of various relationships. In this initial data curation stage, all decipherable relationships were converted to a uniform format (where "C" equals cousin and "R" equals removed). Submissions with indecipherable relationships were eliminated. Submissions with obvious data entry errors were also eliminated, such as those where the longest segment was longer than the total shared cM, or where there was text in the cM field instead of a number.

This initial data curation eliminated a total of 171 data submissions (1.7%), bringing the total to 9,720 data points used for statistical analysis.

Data Analysis

A total of 34 relationships ranging from Parent/Child to 8C were analyzed individually. The total number of submissions for each relationship varied, with a low of six for great-great-aunt/uncle and a high of 889 for aunt/uncle/niece/nephew. A total of 17 of the 34 relationships (52.9%) had 100 or more submissions, and 9 of 34 relationships (26.5%) had 500 or more submissions. See Table 1, below.

A box plot was created for each relationship, and extreme outliers were identified (Q1 - 3*IQR or Q3 + 3*IQR) and removed from the data. Although this approach for removing outliers is widely accepted, outliers should only be removed if there is sufficient justification. A concern with a previous version of data published from the Shared cM Project, in which outliers remained, was that there were extreme minimums and maximums which did not correlate to values actually seen by genetic genealogists and were highly unlikely based on current understanding of genetics. For example, the minimum for Aunt/Uncle/Niece/Nephew was 121 cM when outliers were included. Since the expected amount for this relationship is 1750 cM, the value of 121 cM is most likely due to either an incorrect relationship or a data entry error. Genealogists relying on a range of Aunt/Uncle/Niece/Nephew as low as 121 cM could make incorrect conclusions. Accordingly, there was sufficient justification to remove outliers from the data. Although removing outliers has a significant impact on the data, it arguably results in a dataset with greater reliability.

Table 1. Number of Submissions for Each Relationship Following Outlier Removal

Relationship	Number of Submissions
Aunt/Uncle/Niece/Nephew	889
2C1R	884
1C	869
2C	867
1C1R	839
3C	794
Siblings	789
Parent/Child	758
3C1R	547
Grandparent/Grandchild	281
4C	221
1C2R	193
Half Siblings	187
2C2R	172
4C1R	164
Great Aunt/Uncle	158
3C2R	114
5C	99
5C1R	82
Half Aunt/Uncle	80
Half 2C	51
6C1R	46
7C1R	42
Half 2C1R	40
6C	38
4C2R	34
Half 1C1R	32
Great Grandparent/Grandchild	29
5C2R	25
8C	25
Half 1C	23
6C2R	20
7C	19
Great Great Aunt/Uncle	6
Total	**9,417**

Following outlier removal, the dataset contained 9,417 submissions (96.9% of the total 9,720 submissions). The minimum, average, and maximum values of the remaining data points were identified for each relationship using standard methodology. See, Table 2.

Table 2. Minimum, Average, and Maximum Values for Each Relationship

Relationship	#	Min	Average	Max
Parent/Child	758	3266	3471	3720
Siblings	789	2150	2600	3070
Half Siblings	187	1320	1753	2134
Grandparent/Grandchild	281	1272	1765	2365
Great Grandparent/ Grandchild	29	547	850	1110
Aunt/Uncle/Niece/ Nephew	889	1301	1744	2193
Half Aunt/Uncle	80	540	863	1172
Great Aunt/Uncle	158	521	857	1138
Great Great Aunt/Uncle	6	214	434	580
1C	869	533	880	1379
Half 1C	23	236	554	704
1C1R	839	115	433	753
Half 1C1R	32	78	187	253
1C2R	193	27	235	413
2C	867	43	238	504
Half 2C	51	0	123	245
2C1R	884	0	129	325
Half 2C1R	40	0	73	196
2C2R	172	0	81	201
3C	794	0	79	198
3C1R	547	0	56	156
3C2R	114	0	36	82
4C	221	0	31	90
4C1R	164	0	20	57
4C2R	34	0	14	27
5C	99	0	17	42
5C1R	82	0	14	41
5C2R	25	0	16	41
6C	38	0	9	21
6C1R	46	0	9	19
6C2R	20	0	11	29
7C	19	0	7	10
7C1R	42	0	7	14
8C	25	0	9	16

For relationships where the minimum value was 0 cM shared, the averages were calculated only for cM amounts greater than 0 cM. Accordingly, these averages represent the average only for cousins actually sharing a detectable amount of DNA.

A histogram was created relationships with 100 or more submissions (with the exception of half aunt/uncle, which had 80 submissions). The histograms were created in Excel using the data for each relationship with outliers removed. An example of the histogram for Aunt/Uncle/Niece/Nephew is show below:

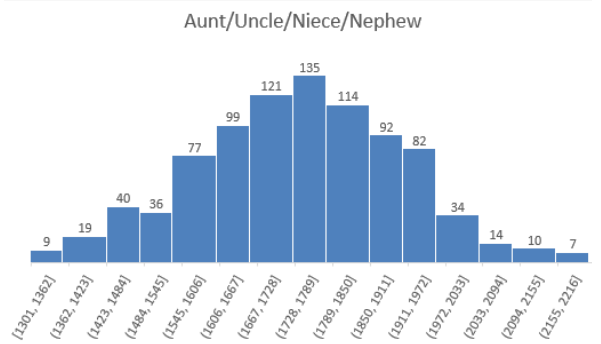


Figure 3. Histogram showing the distribution of shared DNA (in centimorgans) between pairs of people who are aunt/uncle/niece/nephew to one another.

Comparison to Other Data

A comparison of the average values for the data to the ISOGG Expected Shared DNA table (http://www.isogg.org/wiki/Autosomal_DNA_statistics) reveals that the averages are very similar to those expected.[1] However, as discussed above, the expected values do not provide any insight into the ranges of values observed by test-takers.

Table 3. Comparison of the average values in this study to the ISOGG Expected Shared DNA table (http://www.isogg.org/wiki/ Autosomal_DNA_statistics).

Relationship	Shared cM Project (Average)	ISOGG Table (Expected)
Parent/Child	3471	3400
Siblings	2600	2550
Half Siblings	1753	1700
Grandparent/Grandchild	1765	1700
Aunt/Uncle/Niece/Nephew	1744	1700
Half Aunt/Uncle	863	850
1C	880	850
Half 1C	554	425
1C1R	433	425
2C	238	213
Half 2C	123	106
2C1R	129	106
3C	79	53

[1] A comparison of the average values for the data to the ISOGG Expected Shared DNA table (http://www.isogg.org/wiki/ Autosomal_DNA_statistics) reveals that the averages are very similar to those expected (Table 3).

Future Directions

There are many possible avenues for future research and analysis using the Shared cM Project dataset, which continues to grow. Among these possibilities are the following:

Source Breakdown – One of the variables reported for each submission was the source of the information (23andMe, AncestryDNA, Family Tree DNA, or GEDmatch). Determining minimum, maximum, and average values for each testing company and third-party tool individually may reveal important differences.

Endogamy Breakdown – Another variable reported for each submission was whether there was any known endogamy or pedigree collapse in the family tree that could affect the amount of DNA shared by the two test-takers. The current analysis used submissions regardless of their endogamy status. Known endogamy or pedigree collapse is hypothesized to increase the average amount of DNA shared by test-takers compared to those without known endogamy or pedigree collapse.

Group by Clusters – Grouping the data by relationships that share comparable amounts of DNA (rather than by individual relationships) before performing the data analysis may be beneficial. Each cluster will have significantly more submissions than individual relationships. Potential clusters are shown in Table 4.

Table 4. Potential clusters of relationships sharing similar amounts of DNA.

Cluster	Included Relationships
1	Parent/Child
2	Siblings
3	Half Siblings, Grandparent/Grandchild, Aunt/Uncle/Niece/Nephew
4	Great Grandparent/Grandchild, Half Aunt/Uncle/Niece/Nephew, Great Aunt/Uncle/Niece/Nephew, 1C
5	Half 1C, 1C1R
6	Half 1C1R, 1C2R, 2C
7	Half 2C, 2C1R
8	Half 2C1R, 2C2R, 3C
9	3C1R
10	3C2R, 4C

Larger Datasets – As genealogists continue to test family members, the number of submissions to the Shared cM Project continues to grow. In the future, it will be advantageous to repeat this analysis using a greater number of submissions, especially for relationships that are underrepresented in the present version.

Conclusion

The Shared cM Project offers empirical data on DNA sharing that complement existing theoretical and simulated resources for autosomal genealogy tests. It does so by harnessing the power of citizen scientists to amass sufficient data for analysis. The Project can serve as a model for similar group projects to address questions of importance to the genetic genealogy community.

Conflicts of Interest

Blaine Bettinger is an author, educator, and blogger on topics related to genetic genealogy. As a lawyer, he also represents GEDmatch, Inc. before the U.S. Patent and Trademark Office. He declares no conflicts of interest.

Journal of Genetic Genealogy 8(1):43-46, 2016

Product Review: Genome Mate Pro

Leah Larkin, Ph.D.
http://www.theDNAgeek.com

With the rapid rise in autosomal DNA testing over the past few years, data management has become a serious challenge. For example, as of 6 October, 2016, I had 1,036 matching relatives in FTDNA's Family Finder, 1,611 at 23andMe (of which 380 are "sharing" segment data with me), 2,742 at GedMatch (using Matching Segment Search), and 7,268 at Ancestry.com. Trying to cross-reference which of these relatives have tested at multiple sites, track how they triangulate with one another, maintain research notes on particular individuals, and update the data as new matches arrive was a Sisyphian task for just one tester, much less several.

Enter *Genome Mate Pro* (GMP) by Becky Mason Walker. GMP (https://genomemate.org/) is a utility for managing autosomal DNA segment data (and other genealogical information) on Windows, Max, or Linux operating systems. It replaces the earlier *Genome Mate*. GMP is supported by an extensive User Guide developed by Jim Sipe, videos, and a support group on Facebook (https://www.facebook.com/groups/GenomeMatePro/). Best of all, the program remains supported by donations.

GMP is too powerful and complex a program to describe in complete detail. Instead, I will summarize the main features that might influence a new user to try it by describing the available tabs: Profiles, Chromosomes, Relatives, Ancestors, Segment Map, Options, and Help. This review is also not a "how to" guide. For that, readers are referred to the excellent User Guide. Note that I'm using an option called "Privatize Display for Sharing" to truncate the user names; normally, I can see the full name provided by each tester.

The Profiles Tab

The first step in using GMP is to set up a profile in the Profiles tab (Figure 1) for one or more testers. One person who has tested at multiple sites has a single profile to manage all of their data, and each unique tester has their own "profile". The eight fields in the center collect information to consolidate the test results from different sources. For example, I have tested at 23andMe, FTDNA, and Ancestry and am also at GedMatch. By filling in all eight fields according to the instructions on the right, I can manage all of those results within a single profile for myself in GMP. (Of course, a profile still works with data from only one source.)

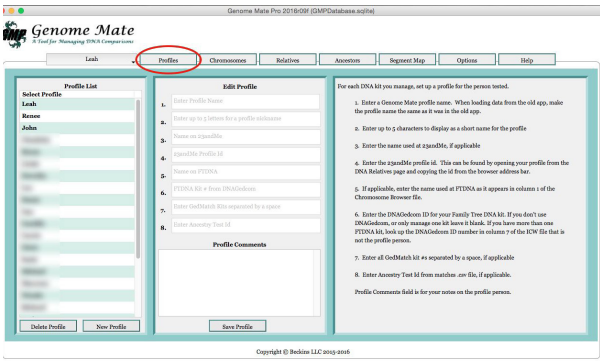


Figure 1. The Profiles tab. Some profile names have been blurred for privacy.

Once you set up a profile, you can import the data from Genome Mate (the predecessor to GMP), 23andMe, 529andYou, FTDNA, GedMatch, DNAGedcom, or AncestryDNA using a set of templates in the Options tab. The templates allow for easy transitions when a data source changes its export format. Advanced users can also create their own templates to import data from a customized spreadsheet.

You can switch easily between profiles in any of the other tabs by using the pulldown at the top left of the GMP window.

The Chromosomes Tab

The Chromosomes tab is where you will do most of your work. It has several powerful features that are numbered in Figure 2 for easy reference. A pulldown (1) lets you change profiles, another pulldown (2) switches between chromosomes (chr 6 is shown here), and a set of filters (3) lets you select which data you want to see. For

example, you might want to see only maternal matches, only segments above a certain threshold (set in the Options tab), or only matches from 23andMe. The bulk of the screen is dedicated to the DNA matches of the profile person. The Profile and Chr (chromosome) columns reflect the pulldown settings. The other columns (4) are all sortable. Note that the relative names are color coded (5) by the source of the segment data. Default colors are lilac for GedMatch, maroon for FTDNA, green for 23andMe, and black for data that has been merged from two or more sources. Names in bold have had a most recent common ancestor (MRCA) designated for that segment (see Relatives Tab). Segment start and stop points (in bp), cM, and SNPs are imported from the original source, while Side and Group are assigned by the user. Note that some of my matching segments are on my maternal side (M), others on my paternal side (P), and the remainder have not been assigned. ("B" for both and "I" for IBS are also options.) Bryan is my mother's first cousin; I designated his group MGF, because he is related to me through my maternal grandfather.

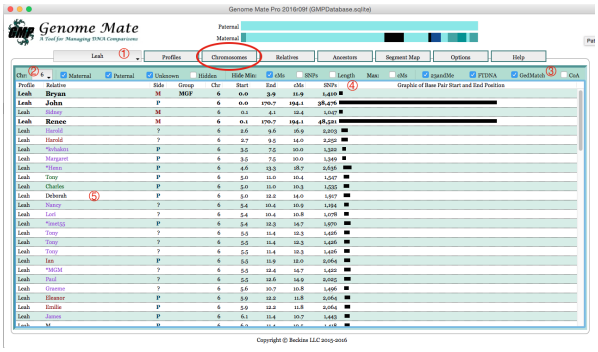


Figure 2. The Chromosomes tab. The names of the matches have been truncated for privacy. Refer to the text for an explanation of the red numbers.

One of the most labor-saving features of GMP is the ease of marking possible triangulations. Right-click on a relative's name for a suite of options. In this case, I chose my mother, Renée. Select "Show possible triangulations", and GMP will filter the list of matches to include only those who share that particular segment with both me and my mother. Right-click again on that relative in the filtered list and select "Mark shown DNA segments" to have all of those potential triangulations marked with the same side and group as the selected relative. Because I have profiles for both of my parents in GMP, I was able to assign all but a few of my matches on chr 6 to either my maternal or paternal side with just a few clicks. Similarly, I could easily mark everyone who triangulates with cousin Bryan as MGF.

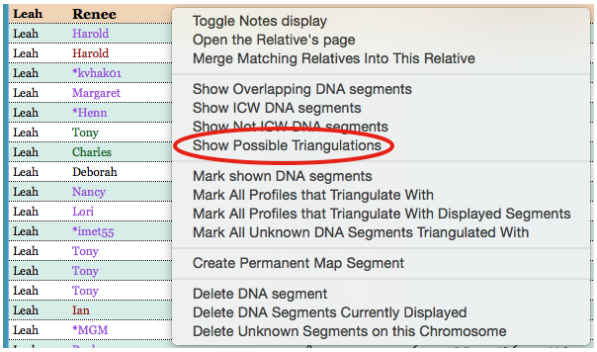


Figure 3. Actions available in the Chromosomes tab.

How does GMP triangulate? It accepts imported triangulation data from the GedMatch Tier 1 tool or from 529andYou, but it can also find segments that match two (or more) profiles. This means that "Show possible triangulations" works even for FTDNA data and for GedMatch users who don't have Tier 1 access, as long as you've set up multiple profiles in GedMatch.

If you hover your cursor over the segment map at the top, GMP will tell you the MRCA to which a particular segment is assigned. If you click on the segment, GMP will filter to only those matching segments that can be attributed to that MRCA. Information from this segment map can also be visualized in as a full genome map in the Segment Map tab. MRCAs are assigned in the Relatives tab.

The Relatives Tab

The Relatives tab compiles information about individual matches. It has four subtabs: About, Family Comparison, DNA Comparison, and Merge. The About subtab contains basic imported data about your match: name, contact information, haplogroups, links to their profiles at the testing companies, and heritage (ethnicity). It also has fields for your research and MRCA notes. The

Family Comparison subtab lists family locations, surnames, ancestors, ancestor details, surnames in common with the profile person, and possible connections. The DNA Comparison tab lists all of the matching segments between that relative and each of your profiles, as well as overlapping segments, in-common-withs (ICWs), and possible triangulated segments. You can also assign a segment to a particular MRCA (or couple) if you have imported a gedcom for the profile person. Gedcoms are imported in the Ancestors tab.

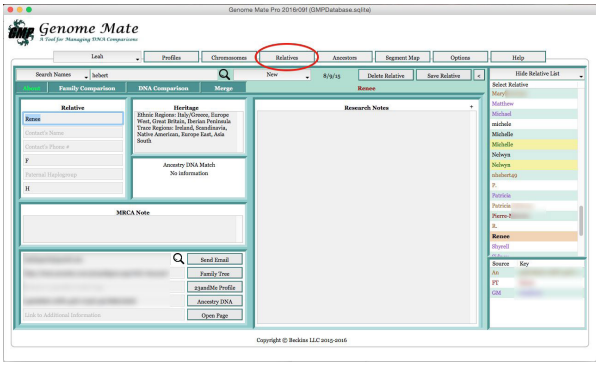


Figure 4. The Relatives tab.

In Figure 2, there were two Harolds in my chromosome browser, one imported from GedMatch (in lilac text) and the other from FTDNA (maroon). They are the same person. The Merge subtab allows me to see the details of the two records side-by-side and easily merge them. All of the identifiers (GedMatch #, FTDNA) will be stored in the merged record. Harold's name will then be shown in black text in the Chromosomes tab.

The Ancestors Tab

GMP allows you to import a gedcom of direct ancestors for each profile. The names in the list on the left are color-coded by maternal/paternal, and radio buttons at the bottom let you filter by side and even by the ancestors who may have contributed X-chromosomal DNA. Gedcom information allows you to assign MRCAs to segments in the Relatives tab. Those assigned segments can then be visualized in the Chromosomes and Segments Map tabs.

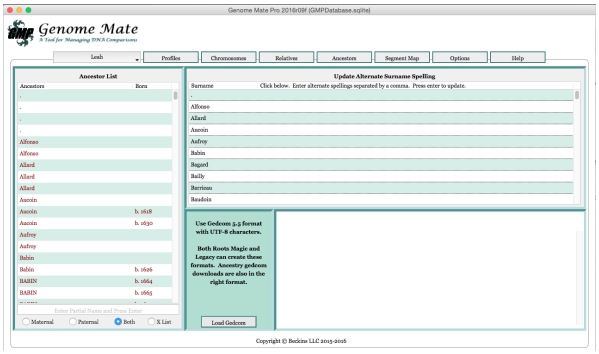


Figure 5. The Ancestors tab.

The Segment Map Tab

At the top of the Chromosomes tab (Figure 2) is a map for that particular chromosome showing segments that have been assigned to MRCAs. The Segment Map tab allows you to view all 23 chromosomes together.

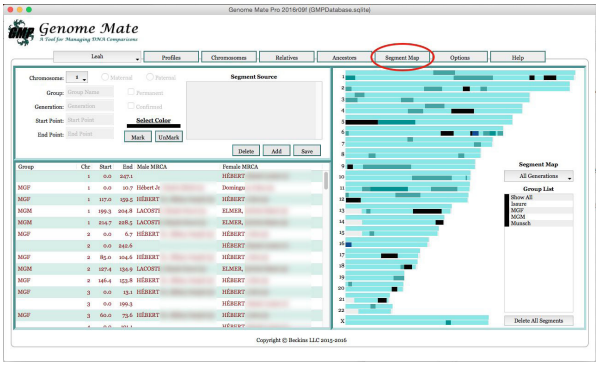


Figure 6. The Segment Map tab.

The Options Tab

The Options tab has a suite of settings to let you customize GMP to your preferences, including setting a threshold segment size for imports and for the chromosome browser display. (This is where I set "Privatize Display for Sharing".) There is also a customizable email template that can auto-fill the match's name and GedMatch number.

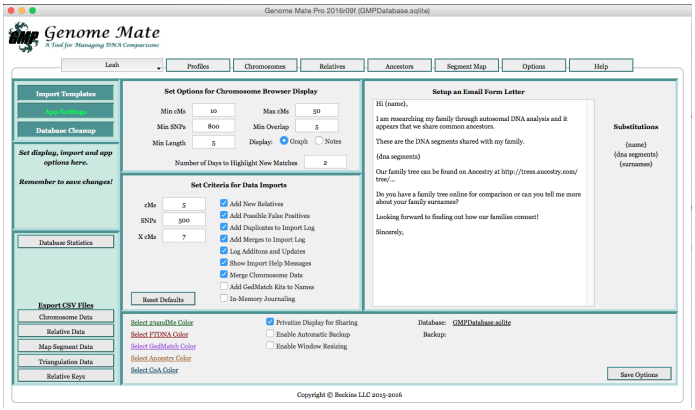


Figure 7. The Options tab.

The Help Tab

The Help tab provides a basic overview for getting started on the right, links to useful resources in the middle, and a list of the most common problems in the "Developer's Comments" on the left.

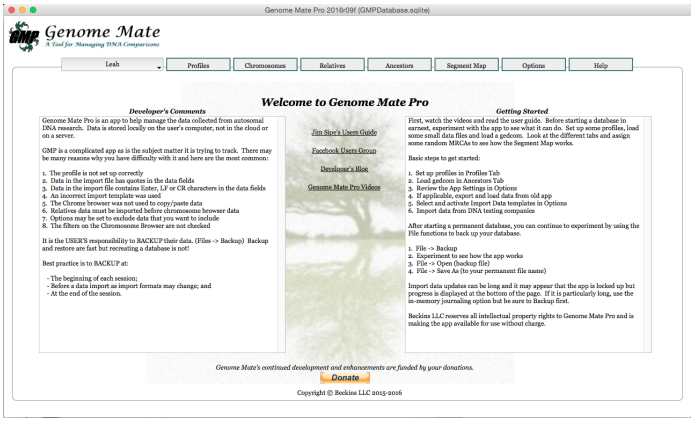


Figure 8. The Help tab.

Drawbacks

The many advantages of GMP should be apparent in the sections above. The biggest drawback is speed; some data imports can take hours, and the program often lags when switching between tabs. GMP also has a few minor inconveniences. Setting up the profiles correctly can also be frustrating, because a new user may not realize the profile information was incorrect until after the import fails.

Finally, caution must be used with the "Show Possible Triangulations" feature because it does not impose a minimum cM threshold. Thus, it can sometimes indicate a triangulation when two profile people overlap on that segment by only a few cM. Users from endogamous background should be especially alert.

Summary

GMP is unequalled in the ease with which a genetic genealogist can manage the complex data associated with DNA matches. The elegant design allows various activities to be separated logically into different tabs while synchronizing information among them. Additionally, GMP is free (although I encourage users to donate if they find it useful). GMP is not for the novice genetic genealogist, but it will prove invaluable to those who are serious about working with segment data.

Conflicts of Interest

The author was part of an early alpha testing group for GMP but has no financial or personal ties to the developer, Becky Mason Walker, or to Jim Sipe, who wrote the User Guide. She is the Editor of the *Journal of Genetic Genealogy*; this manuscript was handled by another member of the Advisory Board. She declares no conflicts of interest.

Book Review

Genetic Genealogy In Practice

By Blaine T. Bettinger and Debbie Parker Wayne
Published by National Genealogical Society; 3108 Columbia Pike, Suite 300, Arlington, VA 22204-4304; www.ngsgenealogy.org; 2016. 196 pp. Appendixes, illustrations, exercises. Paperback. $29.95
ISBN 978-1-935815-22-8.

Two pioneers of genetic genealogy education, Blaine T. Bettinger, Ph.D., J.D., and Debbie Parker Wayne, CG, teamed up to create a resource that provides in-depth information and skill-building exercises for those interested in the intersection of DNA and genealogy.

The book is the vision of Debbie Parker Wayne, who was a driving force behind genetic genealogy education at the Genealogical Research Institute of Pittsburgh, Pennsylvania, USA; Salt Lake Institute of Genealogy, Utah, USA; and the Institute for Genealogical and Historical Research, Birmingham, Alabama, USA. Bettinger began lecturing on DNA for genealogy in 2009, around the same time as Wayne. Bettinger serves as the chairman of the Genetic Genealogy Standards Committee. Bettinger and Wayne's experiences in the genealogical classroom have given them insight into the "continuing need for education and hands-on exercises that help genealogists understand the benefits and limitations of DNA testing" (p. 1).

Topics include basic genetics, standards and ethics, Y-DNA, mitochondrial DNA, autosomal DNA, and X-DNA in addition to chapters on interpreting DNA testing in family studies and incorporating DNA evidence in a written conclusion. There is a section on recombination, a topic that Bettinger feels is critical for those participating in genetic genealogy. Information about the direct-to-consumer genetic testing companies and third-party tools is interwoven throughout the chapters.

The major strength of the publication is in the workbook format. The exercises urge the reader to think about the application of DNA to real life situations. They illuminate circumstances that researchers might not otherwise consider and address how to select the best candidates for testing. The answers to the exercises provide thoughtful, concise, and accurate explanations, often presenting multiple points of information in the answer to a single question.

Genetic genealogists who are interested in documenting their research will appreciate Chapter 8, "Incorporating DNA Evidence into the Written Conclusion." This chapter covers standards, privacy concerns, sharing DNA results, and citing DNA test results. There is a basic discussion of proof argument elements and process as well as suggested topics to be included in an analysis.

The appendices include a glossary and a reading and source list. Charts, tables, and illustrations present visual learning opportunities. Source citations are included.

While *Genetic Genealogy in Practice* is a useful guide for both the beginner and the experienced genetic genealogist, researchers with no previous genetic genealogy experience may wish to complement it with *The Family Tree Guide to DNA Testing and Genetic Genealogy* by Blaine T. Bettinger, which goes into greater detail about some concepts.

Bettinger and Wayne have met the challenge of providing a hands-on learning experience for genealogists that will remain useful until genetic genealogy testing and analysis endures a major change.

Jennifer Armstrong Zinck
North Granby, Connecticut
jenzinck@gmail.com