Matching Bettinger: Small Segments, A Big Problem

James Sims, PhD

Published: October 31, 2017

Background

On October 16, 2017, Blaine T. Bettinger, PhD, JD[1] issued the following challenge to the members of the Genetic Genealogy Tips & Techniques Facebook group[2], which he founded earlier that year and that he co-administers. “Another exercise to emphasize the futility of small segments. Here’s my GEDmatch Kit #: A812216. Go to GEDmatch and do a”‘One-to-one’ Comparison" to with my Kit #, BUT lower the SNP threshold to 100 SNPs and the match threshold to 3 cM (see the first comment for instructions). Do you “match” me at this level?" In a second challenge Bettinger wrote, “Does PHASING help? Here’s my phased maternal GEDmatch Kit #: PA812216M1. Go to GEDmatch and do a”‘One-to-one’ Comparison" to with my Kit #, BUT lower the SNP threshold to 100 SNPs and the match threshold to 3 cM (see the first comment for instructions). Do you “match” me at this level? [Inspired by Ann Turner, thank you!]"

Introduction

The field of genetic genealogy is a twenty-first century development following the publication of the draft sequence of the human genome in 2001[3]. Today, tens of thousands of genealogists, among the more than seven million people[4] who have taken a DNA test for ancestry around the world, turn to DNA analysis in the hope of breaking through some of their genealogy brick walls. One use of DNA for genealogy is cousin fishing, which is an attempt to identify new cousins by DNA who may know more about an ancestral family than the test taker does. Autosomal DNA (atDNA) analysis, the most popular type of DNA analysis for ancestry today, is most powerful when trying to detect immediate and close family members and first, second, third and fourth cousins. The atDNA matching technique loses power to detect known more distant cousins as a consequence of the biological mechanism by which atDNA is passed down from generation to generation, and the statistical algorithms used to call matches between individuals[5].

A second use of atDNA matching is to perform genealogical hypothesis testing. In hypothesis testing, the binary situation of being a DNA match, or not matching, is proposed to be evidence of common biological descent (or lack thereof) in a specific genealogical time frame. The amount of shared atDNA expressed in units of cM often plays an important role in hypothesis testing. When using atDNA analysis for hypothesis testing, genetic genealogists must be aware of the limitations of the technique.

Phasing, as the term is used in genetic genealogy, refers to the computational procedure of assigning DNA segments in an atDNA analysis to the chromosomes of maternal and paternal origin. There are several variations of the phasing technique, the most powerful method being the comparison of atDNA of two parents and a biological child[6]. The purpose of phasing is to correctly assign each DNA segment to a paternally

[1] Bettinger, Blaine, PhD, JD, author of The Genetic Genealogist blog, https://thegeneticgenealogist.com/

[2] Genetic Genealogy Tips & Techniques is a closed Facebook group with over 20,000 members as of the writing of this article. All posts must be about some aspect of using DNA for a genealogical purpose. The group has members with all levels of experience in genetic genealogy.

[3] Venter, J. C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., et al. The sequence of the human genome, Science 291: 1304-1351, 2001

[4] Larkin, Leah PhD, Genealogical Database Sizes—August 2017 Update, http://thednageek.com/genealogical-database-sizes-august-update/, accessed October 29, 2017

[5] Cousin statistics, International Society of Genetic Genealogy Wiki, https://isogg.org/wiki/Cousin_statistics, accessed October 29, 2017

[6] Phasing, International Society of Genetic Genealogy Wiki, https://isogg.org/wiki/Phasing accessed October 29, 2017

inherited or a maternally inherited chromosome, and to reduce the rate of false positive matches when comparing DNA between individuals.

Novice genetic genealogy hobbyists and others are often referred to powerful online software tools, such as those available at GEDmatch[7], which provide tools like a chromosome browser, something that is not available to customers of certain atDNA testing companies. Also, hobbyists use GEDmatch to find DNA matches to individuals who tested at a different company. Unlike atDNA testing vendors, which choose the matching threshold criteria[8] for their customers, users of tools at GEDmatch are free to reset the matching criteria to their liking. With little knowledge of the consequences when lowering the matching thresholds, it is possible for genetic genealogists to falsely discover new DNA matches. Once a false discovery is made, the use of that match for cousin fishing and for hypothesis testing may lead to a considerable amount of wasted time, and erroneous conclusions, respectively. This article is intended to provide an example why DNA testing companies choose certain matching criteria, and why lowering match criteria below vendor set thresholds can lead to additional false discoveries.

Results

Table 1 shows the results of attempts to match the author's atDNA kits at several vendors to kits belonging to Blaine Bettinger at those same vendors. The author searched his match lists for kits with an associated Bettinger surname. These attempts to find matches serve as control experiments for the experiments described later in this article. As shown in Table 1, the author was not a DNA match to Blaine Bettinger according to three DNA testing companies based in the United States. If there was a match to a kit with Bettinger's name, the Match variable in Table 1 was scored as T for true, otherwise F for false. In the cases in which Match was false, the values of the other variables were reported as NA, meaning not available.

Table 1: Matching Controls at Commercial Vendors

Vendor	Match	Total cM Shared	Number of Segments Shared	Largest cM Value
23AndMe[9]	F	NA	NA	NA
AncestryDNA[10]	F	NA	NA	NA
Family Tree DNA[11]	F	NA	NA	NA

In Table 2, the results of attempts by the author to match his kit at GEDmatch to Blaine Bettinger's kits at GEDmatch are shown. In the case of Control 1 and Control 2 shown in Table 2, the cM threshold and SNP count threshold was set to the default values for the One-to-One match tool at GEDmatch. Control 1 shows the results of attempting to match the author's kit to Bettinger's unphased kit. Control 2 shows the attempt to use the default values at GEDmatch to match Bettinger's phased kit. In both cases, the author was not a match.

Experiments A-E were conducted using Bettinger's phased kit. In Exp. A shown in Table 2, the SNP count threshold was lowered to 250 from the default value of 500, and again there were no matches between the author's kit and Bettinger's phased kit. In Exp. B, the cM threshold was lowered to 3.0, which was the value mentioned in Bettinger's challenges, however in this experiment, the SNP count threshold was set at 250. No match was detected. In Exp. C, the SNP count threshold was lowered to 200 from 250, and the result was a match involving one DNA segment with a cM value of 3.9. In Exp. D, the SNP count threshold was lowered to 150, and the result was three matching DNA segments with 10.3 cM total shared DNA, and the largest cM value for a segment was 3.9. In Exp. E, the SNP count threshold was lowered to 100, which was

[7] Gedmatch web site, https://gedmatch.com

[8] Criteria for matching segments, Autosomal DNA testing comparison chart, International Society of Genetic Genealogy Wiki, https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart accessed October 29, 2017

[9] 23AndMe web site, https://www.23andme.com/

[10] AncestryDNA web site, https://www.ancestry.com/dna/](https://www.ancestry.com/dna/)

[11] Family Tree DNA web site, https://www.familytreedna.com/

Bettinger's challenge setting, and the cM threshold was 3.0, also Bettinger's challenge setting. With these settings of the One-to-One match tool at GEDmatch, there were five matching segments with a total shared cM value of 17.2 and the largest cM value was 3.9. In Exp. F, Bettinger's challenge settings were used to compare the author's kit to Bettinger's unphased kit. The results were 31 matching segments, the largest of which was 4.6 cM, and the total shared cM value was 109.1.

Table 2: Effect of Changing Matching Criteria at GEDmatch

Obv.	Match	Phase	cM Thrs.	SNP count	Total cM Shared	Number of Segments Shared	Largest cM Value
Ctrl 1	F	F	7.0	500	NA	NA	NA
Ctrl 2	F	T	7.0	500	NA	NA	NA
Exp. A	F	T	7.0	250	NA	NA	NA
Exp. B	F	T	3.0	250	NA	NA	NA
Exp. C	T	T	3.0	200	3.9	1	3.9
Exp. D	T	T	3.0	150	10.3	3	3.9
Exp. E	T	T	3.0	100	17.2	5	3.9
Exp. F	T	F	3.0	100	109.1	31	4.6

In Table 2, the letters T and F in the table indicate a variable was either true or false, respectively. In the case of the Phase variable, this refers to Bettinger's phased kit shown as a T, or his unphased kit shown as F. Abbreviations: Ctrl, control; Obv., observation; Thrs., threshold; SNP, single nucleotide polymorphism.

Discussion

The experiments conducted for this article illustrate the problem users of GEDmatch's One-to-One comparison tool may encounter when they arbitrarily lower the matching criteria. The author had no genealogical reason to think he was related to Bettinger in a meaningful genealogical time frame, say within the last 20 generations. The results of searching for Bettinger's kit among the DNA matches the author has at three DNA testing companies in the United States showed the author was not a DNA match to Bettinger. When the matching criteria are systematically lowered for the author's unphased kit at GEDmatch, he was eventually able to generate more than 100 cM of matching DNA between his kit and that of Bettinger's unphased kit. The largest of these matching segments had a 4.6 cM value. Novice genetic genealogist might conclude this match was significant, when it is clearly not significant. It is either a false positive match in a meaningful genealogy time frame or a rather ancient match. Either way, such matches are unlikely to be useful for genealogy.

Experiments E and F compare the effect of using phased and unphased kits for DNA matching, respectively. The author falsely shared 31 segments with Bettinger when comparing his unphased kit to Bettinger's unphased kit. The effect of making a comparison to Bettinger's phased kit was to reduce the number of falsely discovered matching segments from 31 to 5. Thus, phasing clearly reduced the number of false matching segments, but in the case shown here, phasing did not eliminate the possibility of false/un-useful discoveries at low matching thresholds.

Methods

The author tested atDNA at AncestryDNA in 2014, and the results were reported from AncestryDNA array V1.0, according to the header information in the raw data file. The author's kit at GEDmatch was created previously by uploading the raw autosomal data obtained from the author's AncestryDNA account. GEDmatch experiments were conducted online using the Google Chrome web browser. The results of each run of the One-to-One match tool at GEDmatch were copied and saved to Evernote[12] for archival purposes.

[12] Evernote web site, https://evernote.com/

The results were transcribed by hand for this article. This document was authored in rmarkdown using the free version of RStudio[13], the statistical programming environment called R[14], the knitr package[15] and its dependencies, and MacTEX for conversion of rmarkdown to PDF format[16].

[13] RStudio web site, https://www.rstudio.com/

[14] R Project web site: https://www.r-project.org/

[15] knitr package on CRAN, https://cran.r-project.org/web/packages/knitr/index.html

[16] MacTEX web site, http://www.tug.org/mactex/