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:

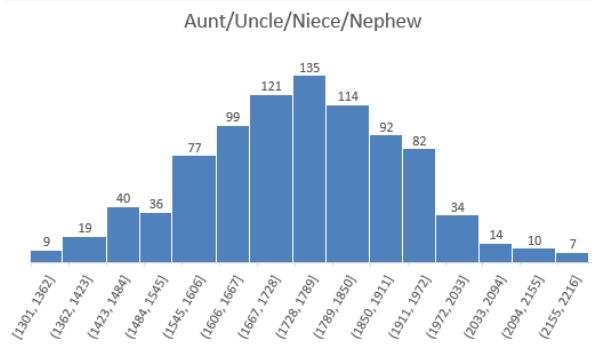

Figure 3. Histogram showing the distribution of shared DNA (in centimorgans) between pairs of people who are aunt/uncle/niece/nephew to one another.

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.