Sharing Distant Autosomal DNA: Low probability is not no probability

by Wesley Johnston (6398)

Abstract

The “Birthday Paradox” befuddles most people: if 23 people are together, the odds are 50% that two will share the same birthday, even though 23 is only 6% of the 365 days in a year. The same Mathematics that underlies the “Birthday Paradox” underlies the reality that far fewer distant cousin DNA kits than expected by most people are needed to provide 99% probability that at least two of those kits will match at 7 cM or more. This primarily benefits DNA projects that include many kits from the same families who are probably related in the 1700s. It is of little benefit to the lone researcher who is not part of a project. The application of this reality in three projects includes one that disproves the assumption that all people with 1600s colonial American ancestors are related to each other due to a highly restricted marriage pool.

Probability of Sharing Distant DNA with a Cousin

We inherit 50% of our autosomal DNA from each parent and roughly 25% of our DNA from each of our four grandparents. Each generation back reduces it by half. So, you have roughly 1.6% of your DNA from each of your 64 4th great grandparents.

Your 4th great grandparent may have many living descendants in your generation (your 5th cousins), but each of them has only about 1.6% of their DNA from that 4th great. And their 1.6% may be completely different DNA from your 1.6%.

So, the odds of you and any one of those 5th cousins sharing the same DNA are small. But they are not as small as 1.6%. At the ISOGG Wiki Cousin Statistics web page, you can find the “probability that two cousins will share enough DNA for the relationship to be detected”.[1]

The table below shows how these numbers (using the 23andMe detection probabilities) determine the number of pairs of testers needed and thus the number of kits needed to provide 99% probability that at least one pair of testers will share detectable DNA from their common ancestor at that cousin level. (See the end section for the details of the Math.)

Cousin Level	Percent Inherited	Detection Probability	Pairs Needed	Kits Needed
1	25.000%	100.000%	1	2
2	12.500%	100.000%	1	2
3	6.250%	89.700%	3	3
4	3.125%	34.900%	11	6
5	1.563%	14.900%	29	9
6	0.781%	4.100%	111	16
7	0.391%	1.100%	417	30
8	0.195%	0.240%	1917	63
9	0.098%	0.060%	7,673	125
10	0.049%	0.002%	230,257	680

The 23andMe web page “DNA Relatives: Detecting Relatives and Predicting Relationships” tells what they detect: “Our simulations have concluded that we can confidently detect related individuals if they have at least one continuous region of matching SNPs … that is longer than our minimum threshold of 7cM … long and at least 700 SNPs.”[2] And on 23andMe’s page “The Probability of Detecting Different Types of Cousins”, they write: “Note that even though there is a relatively low chance of detecting more distant cousins, DNA Relatives will likely find quite a few given the large number of distant cousins that exist.”[3]

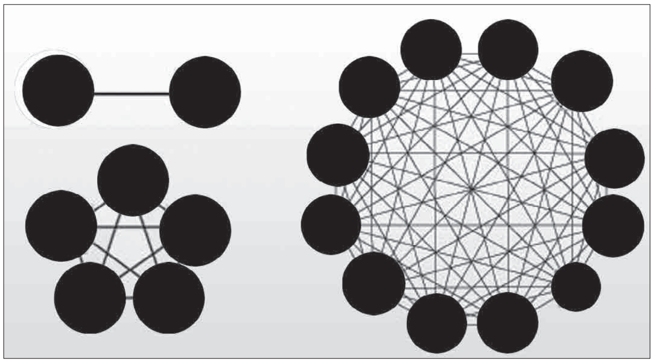
The number of testers needed to find two of them who share enough DNA for their relationship to be detected with 99% probability is surprisingly low. You only need nine 5th cousin descendants of their common 4th great grandparent to have 99% probability that at least two of those 9 kits will share detectable DNA inherited from that 4th great grandparent.

And with sixty-three 8th cousins, you have 99% probability that at least two of them will share detectable DNA inherited from that 7th great grandparent.

Of course, you can also be lucky and find the right pair with fewer kits, but with the number of kits in the table, you have 99% probability of succeeding.

The power of numbers: pairings of individuals

The diagram below shows three different numbers of individuals: 2, 5 and 12. And it shows all of the ways in which those individuals can be paired with each other.



Two individuals make only 1 pair. Five individuals make not 5 pairs but 10 pairs. And 12 individuals make 66 pairs. The number of pairings increases far faster than the number of persons.

The precise number of pairings can be easily calculated. You take the number of people and subtract 1 and then add up all the numbers from 1 to that number. So, for 12 people, subtract 1 to make 11, and then add all the numbers from 1 to 11, making 66. You can more easily calculate this by taking the number of people and multiplying that by one fewer people and then dividing the result by 2. Thus, 12 * 11 / 2 = 66.

For the case of 5th cousins, where you need 29 pairings to have 99% certainty, you really only need DNA for nine of those 5th cousins. This is because nine people connect with each other in 36 different pairs, and you just need 29 pairs for a 5th cousin match to be detectable.

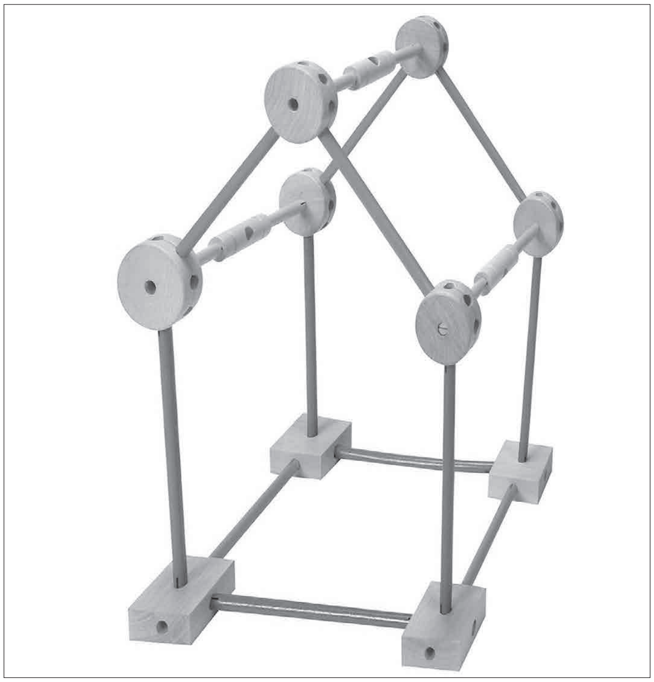
Even for the extreme case of 10th cousins, where you need 230,257 pairs, you would need only 680 10th cousins for there to be 99% probability that at least one pair of them would have inherited enough shared DNA to be detected as 10th cousins.

The power of numbers: project groups

The problem with many statements about using autosomal DNA for finding relatives who connect you to distant ancestors is that they are looking from just one perspective. Yes, it is true that if you have only 16 people of the 6th cousin generation who test, at least two of them will be detectable as 6th cousins. But if I am not one of those two people, it is useless to me, isn't it? Wrong!!

Here is a simple example. Three sisters all test, but only one of them matches a cousin in Germany. And only another one of the three matches a cousin in the USA. And the third sister matches a Canadian cousin who does not match the other two. All three of them are equally related to all those cousins. But if each one of the sisters had used only their own results, then instead of matching three cousins, they would only have matched one.

Think of a set of tinker toys. A rod fits into a hole in a hub piece. You make a building by connecting several rods to a single hub and then connecting each of those rods to a different hub. People are like the hubs, and their DNA matches to other people are like the rods.



You do not build the building expecting that there will be rods connecting every single hub. But that is the expectation that some people have when they try to go it alone with autosomal DNA matching. They want to find the most cousins, but instead of a structure, they wind up with a bunch of pairs or trios of hubs connected in barbells or triangles but not connected much further – not a structure but just a collection of many pieces.

The reality is that you do not connect genetically with every one of your distant cousins. Just as with the three sisters and the tinker toys, you connect to some of them, while some who you connect to also connect to others with whom you do not share DNA. This is how an autosomal DNA project brings the power of numbers of kits to bear on putting distantly-related family members together.

Ancestry's now-abandoned DNA Circles was a good example of this. Autoclustering tools, In Common With tools, Ancestry's Thru Lines and MyHeritage's Theory of Family Relativity give major help in this. Make no mistake about it, you can do powerful analysis with these tools, if you are willing to do the work to verify everything. But all of these tools only let you do analysis with your own kits as the reference point. You need to be able to robustly manage a project where you can see how all the kits – yours and all the others in the structure – relate to each other.

The only tool that allows you to fully manage a project of related kits, is GEDmatch's Tier 1 tag group feature in their Multiple Kit Analysis (MKA), where you can apply the full array of GEDmatch analytical tools to compare all the kits to each other with a single mouse click.

The ultimate power of a project though comes from a committed group of researchers on the specific focus of the group – a surname, a place, a surname in a place, a specific common ancestor or couple. You can do a great deal on your own, but a well-managed project with a dynamic discussion by researchers who share not only DNA with each other but the willingness to dive into the challenges of the documentary and DNA research, robustly enabled by GEDmatch tag groups and MKA – this is a very real power of numbers.

The power of numbers: pairings of individuals in group projects

So, to reap the fullest benefits of your autosomal DNA, you need projects of multiple kits that inter-connect with each other, empowered by GEDmatch multiple kit analysis of your tag group. And it takes relatively few kits or descendants of even a distant ancestor for you to detect distant relationships in at least one pair of those kits.

And the other more-recent relationships that connect members of the project let you build a larger structure from those smaller connections.

The key point is that **LOW PROBABILITY DOES NOT MEAN NO PROBABILITY**. If you have enough (and the number to make "enough" is not that large) descendants of a distant relative who have DNA-tested and you put them together into a project for group analysis, you can have a great deal of success with even distant connections. While it may look horribly small when you see that you only inherit 0.2% of your 7th great grandparent's DNA, the reality is that you need only 30 descendants on different lines to test to have 99% certainty that at least two of them would share DNA detectable DNA inherited from that 7th great grandparent.

Practical Applications

Interpretation of autosomal DNA for distant ancestors has potential problems. Most significantly, autosomal DNA evidence alone is not sufficient to prove a family tree. Documents and other forms of DNA evidence can and should be used to make the case for any interpretation of autosomal DNA.

There are other problems, beyond the scope of this paper. But, just as low probability does not mean no probability, so the problems in interpretation of autosomal DNA do not mean that it cannot be properly interpreted to make part of the case for modern DNA testers matching through DNA they share from distant ancestors. This section examines three such successes.

Dr. Tim Janzen was the first to publish his findings in his presentation "Tracing Ancestral Lines in the 1700s Using DNA" which he has now presented several times. His January 2021 slides are at:

https://drive.google.com/file/d/1TwIiXAKoIB0TpwS6ptLMAXzW6Iz8xOqF/view

In 97 slides, he covers in great detail how he leveraged autosomal DNA, among other forms of evidence, to make solid connections: "Autosomal DNA testing may allow you to break through some genealogical brick walls in the 1800s and possibly the 1700s that exist due to the lack of genealogical records."

Martin McDowell and colleagues from the North of Ireland Family History Society began the Ballycarry DNA Project to examine in depth the DNA of those living in that County Antrim parish who have deep roots there. The project website is at:

https://www.nifhs.org/dna/ballycarry-dna-project/

Martin has made several presentations, one of which is available to Legacy Webinar subscribers at:

https://familytreewebinars.com/download.php?webinar_id=1497

DNA has allowed them to make many connections where no documents exist, including identifying maiden names of wives.

The third project has yet to publish any results publicly. The Loyalist Lake Family History Project has brought together a robust e-mail researcher discussion group who have made many breakthroughs, including using DNA to connect over 100 descendants of common ancestors from the mid-1700s. They have made solid DNA connections where no documents exist and have identified the maiden name of a key ancestor of the largest sub-group.

Being a colonial American project, this project also demonstrates the exception to what has come to be a rule among some genetic genealogists. That rule is that pedigree collapse was the norm among American colonialists because the marriage pool was too small as families remained in the same area for generations. But the Lake family proves the exception to this "rule". And probably the majority of Loyalist families' movements also make the "rule" clearly irrelevant in their case.

The reality is that the Lake family moved often and did not stay in the same marriage pool for long and then went their separate ways to far-flung places. The living descendants of these branches have little worry about pedigree collapse or about the confirmation bias of mistaking the triangulation of kits from five different lines for connecting on the wrong ancestral line when there is only one common ancestral line of the five kits. So, the colonial marriage pool "rule" simply does not fit with the reality of the Lake family – and probably most Loyalist families.

The common theme in the success of all three of these projects is that with sufficient numbers of testers gathered into a unified and focused project, the reality of the probabilities in the tables of this paper have been realized.

Low probability does not mean no probability. And the existence of numerous challenges with interpretation of autosomal DNA for distant ancestors does not mean it is impossible to do proper interpretation and succeed in connections of testers whose common ancestors lived in the 1700s. It is very difficult work to do accurately, which is why so little has been published about it. But it is happening, as will become obvious over the course of time because the reality of the underlying Math in the probabilities is far better than most people are aware: you really do need far fewer kits than most people would think.

The Autosomal DNA Analog of the Birthday Paradox

This is very much the autosomal DNA analog of the counter-intuitive Birthday Paradox. In the birthday paradox, if you have 23 people in a room, the odds are 50-50 that at least one pair of those people will have their birthday on the same day. While 23 is only 6% of the 365 days in the year, it is enough people to give 50% probability of a match.

And if you have just 50 people, it is very nearly certain that at least two of them will share the same birthday. The formula for the probability is:

$$p(n) = 1 - (364/365) \wedge ((n * (n - 1))/2) \qquad (1)$$

where n is the number of people and p(n) is the probability that at least one pair will share the same birthday. It works by subtracting from certainty (1 = 100%) the probability that NO pair share a birthday. (See https://youtu.be/Jn2s1BSMQyM for an excellent explanation of the birthday paradox.)

It all goes back to the complexity diagram in the first part of this paper: the number of pairs grows much faster than the number of people. So, you do not really need huge numbers of people to detect DNA connections with even distant cousins.

The Math: Low Probability is not No Probability

In this paper, I aim to share understanding, without going into the Math. But for those wanting a clearer view of the Mathematics involved, this section is for you.

The following table uses the 23andMe percent probabilities of detectable matches, from the ISOGG Cousin Statistics Wiki web page. The 23andMe probabilities are used as a worst-case scenario, since they are the lowest probabilities on the ISOGG web page.

C	Inherited	Detect	NoDetect	Pairs	Kits
1	25.000%	100.000%	0.000%	1	2
2	12.500%	100.000%	0.000%	1	2
3	6.250%	89.700%	10.300%	3	3
4	3.125%	34.900%	65.100%	11	6
5	1.563%	14.900%	85.100%	29	9
6	0.781%	4.100%	95.900%	111	16
7	0.391%	1.100%	98.900%	417	30
8	0.195%	0.240%	99.760%	1917	63
9	0.098%	0.060%	99.940%	7,673	125
10	0.049%	0.002%	99.998%	230,257	680

“C” is the degree of cousin: 1st, 2nd, etc.

“Inherited” is the expected average percent of one’s own DNA inherited from the common ancestor by two cousins of degree C. “Inherited” does not figure in the calculations of kits needed and is only included for comparison.

$$\text{Inherited} = 0.5^{(C+1)} \qquad (2)$$

“Detect” is the 23andMe probability that two cousins of degree C will share enough DNA for the relationship to be detected.

“NoDetect” is the opposite of “Detect”. It is the probability that two cousins of degree C will NOT share enough DNA for the relationship to be detected.

$$\text{NoDetect} = 100\% - \text{Detect} \qquad (3)$$

“Pairs” is the number of pairs of descendants of the common ancestor needed to give 99% probability that at least one of those pairs will be detectable. The formula for the probability that the number of Pairs for cousin level C will have at least one pair whose shared DNA is detectable is:

$$p = 1 - (\text{NoDetect}^{\text{Pairs}}) \qquad (4a)$$

The similarity of this situation and the Birthday Paradox can be seen by comparing this formula to formula 1 above.

To have 99% (= 1.00 -.01) probability of a match, we set

$$(\text{NoDetect}^{\text{Pairs}}) = .01 \qquad (4b)$$

and solve for “Pairs”, using logarithms or natural logarithms and round up to the next whole number:

$$\text{Pairs} = \text{Ceiling}(\ln .01 / \ln \text{NoDetect}) \qquad (4c)$$

“Kits” is the number of kits of descendants on different lines of the common ancestor needed to reach 99% certainty that at least two of them will share detectable DNA from the common ancestor. If you have some number of Kits, then the number of Pairs that those Kits contain is:

$$\text{Pairs} = (\text{Kits} * (\text{Kits} - 1)) / 2 \qquad (5a)$$

So, to find the number of “Kits” needed to make the necessary number of pairs to give 99% probability of a pair that match at cousin level “C”, we use the quadratic formula to solve for the positive value of “Kits” and round up to the next whole number:

$$\text{Kits}^2 - \text{Kits} - (2 * \text{Pairs}) = 0 \qquad (5b)$$

$$\text{Kits} = \text{Ceiling}(1 + \text{sqrt}((1+(8 * \text{Pairs}))/2))) \qquad (5c)$$

The Ceiling is not appropriate for C<4, so that the values for C<4 in the table are filled in.

Conclusion

Low probability does not mean no probability. It means that you need to have enough kits to deal with the low probability. And the number of kits needed for 99% probability of at least one detectable matching pair is smaller than most people might think.

References

1. https://isogg.org/wiki/Cousin_statistics
2. https://customercare.23andme.com/hc/en-us/articles/212170958-DNA-Relatives-Detecting-Relatives-and-Predicting-Relationships#detecting_a_match
3. https://customercare.23andme.com/hc/en-us/articles/212861317-The-probability-of-detecting-different-types-of-cousins

Wesley is studying the surname Butson and can be contacted at wesley.johnston@one-name.org. Wesley’s registered website can be found at www.wwjohnston.net/famhist/early-butson.htm and his DNA project website at www.familytreedna.com/groups/butson/about.

Instructions for Contributors

We welcome articles, photographs, letters, and news from members.

Please send your submissions to the editor at:

editor@one-name.org

The deadline for the following editions are:

- 15 February
- 15 May
- 15 August
- 1 November

Please note that the Editor reserves the right to amend an article due to various reasons/restrictions and cannot guarantee which edition submissions will appear as this is due to space limitations along with ensuring diversity of content.