Journal: www.jogg.info
Originally Published: Volume 7, Number 1 (Fall 2011)
Reference Number: 71.002

'SATIABLE CURIOSITY: IDENTITY CRISIS: IDENTICAL BY STATE OR IDENTICAL BY DESCENT?

Author(s): Ann Turner

‘Satiable Curiosity

Identity Crisis:
Identical by State or Identical by Descent?

‘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.**

The previous ‘Satiable Curiosity column “Up Hill and Down Dale in the Genomic Landscape”[1] began with this introduction:

> *More genetic genealogy companies are offering genome-wide tests. The latest entrant is Family Tree DNA’s Family Finder,* ***[2]*** *which uses about 500,000 autosomal markers to identify people who share enough DNA to be cousins of some degree. The basic premise of this and similar tests, such as Relative Finder from 23andMe, is that long matching segments are evidence of recent relationships.*
>
> *The Family Finder report includes segments that are too short to be stand-alone evidence for recent relationships, but they may nonetheless repay close scrutiny for insight into the way DNA patterns are distributed within a lineage or within populations.*

I solicited contributions of data with the notion that we might identify short segments common to many people in a population. However, shortly after that column was published, Family Tree DNA (FTDNA) switched to a different chip with about 700,000 markers called SNPs (Single Nucleotide Polymorphisms), so I was unable to collect enough data to draw any useful conclusions.

However, short segments within a lineage are still worth studying, and this column proposes a new quandary. How short is short? At what point can we be confident that a segment is Identical by Descent (IBD) and not just similar due to coincidence (Identical by State, IBS)? This dilemma arises because we are often limited to studying genotypes, where the two alternative versions (alleles) of a marker can not be assigned to the parent who passed them on. All we know is that two alleles are present. Phasing (the process of assigning alleles to the proper parent) is readily accomplished if all three parties, the father-mother-child trio, have been tested.[3] The child’s data can then be presented as two haplotypes (a set of neighboring alleles occurring on the same chromosome and inherited together as a package). But otherwise, the alleles are listed in an arbitrary order: a genotype with the two bases adenine and guanine is always listed as AG but never GA when the raw data is downloaded.

The method for determining a match looks for a long consecutive run of “half-identical” SNPs, where at least one allele in one party matches at least one allele in the other party. A heterozygous result, for instance AG, will be half-identical to all possible genotypes (AA or AG or GG), and it will thus be a half-identical match with nearly everyone in the world.[4] The boundaries of the run are set by the presence of opposite homozygotes, such as AA and GG. These opposite homozygotes crop up often enough to make it difficult to achieve a long consecutive run, unless the two people are IBD.[5]

Hypothetical Case A illustrates how this would appear in a child where paternal and maternal haplotypes have been determined. The father is responsible for the match, since every allele can be accounted for in the paternal haplotype. If the father was tested, he would see the same cousin in his list of matches, and the

segment in the child is IBD. Just by coincidence, the mother also matches some alleles, which would be IBS, as evidenced by the gaps.

Case A – Identical by Descent

maternal allele	paternal allele	child's genotype	putative cousin's genotype	
A	A	AA	GG	opposite homozygotes not part of the run of half-identical SNPs
A	A	AA	AG	cousin is universal match
A	G	AG	AG	both are universal matches
A	G	AG	GG	child is universal match, but the allele came from the paternal side
G	A	AG	AA	child is universal match, but the allele came from the paternal side
G	G	GG	AG	cousin is universal match
A	G	AG	GG	child is universal match, but the allele came from the paternal side
A	G	AG	AG	both are universal matches
A	G	AG	GG	child is universal match, but the allele came from the paternal side
A	A	AA	AA	the A allele is found in 85% of the population, so it's easy to match here
A	G	AG	GG	child is universal match, but the allele came from the paternal side
G	G	GG	AA	opposite homozygotes terminate the run of half-identical SNPs

Case B illustrates a pseudo-segment. The child's genotype and the cousin's genotype are the same, but the haplotypes reveal that the matching allele zig-zags back and forth between the maternal side and the paternal side. The genotype is IBS, and this match would not show up for either parent.

Case B – Identical by State

maternal allele	paternal allele	child's genotype	putative cousin's genotype	
A	A	AA	GG	opposite homozygotes not part of the run of half-identical SNPs
A	A	AA	AG	cousin is universal match
A	G	AG	AG	both are universal matches
G	A	AG	GG	child is universal match, but the allele came from the maternal side
G	A	AG	AA	child is universal match, but the allele came from the paternal side
G	G	GG	AG	cousin is universal match
A	G	AG	GG	child is universal match, but the allele came from the paternal side
A	G	AG	AG	both are universal matches
G	A	AG	GG	child is universal match, but the allele came from the maternal side
A	A	AA	AA	the A allele is found in 85% of the population, so it's easy to match here
A	G	AG	GG	child is universal match, but the match is to the paternal side
G	G	GG	AA	opposite homozygotes terminate the run of half-identical SNPs

Case C is an example of a different haplotype configuration, which would also result in a child's match being absent from either parent's list. The threshold for declaring a match is somewhat arbitrary (ten SNPs in the illustrative cases, hundreds of SNPs in actual testing). The boundaries are not clear-cut, and the true length of the segment may be shorter than the definitive cut-offs of opposite homozygotes. In this case, the child inherited nine consecutive alleles from the paternal side, but he also happened to inherit one allele from the maternal side, which *appeared* to keep the run intact long enough to declare a match.[6]

Case C – fuzzy boundaries

maternal allele	paternal allele	child's genotype	putative cousin's genotype	
A	A	AA	GG	opposite homozygotes not part of the run of half-identical SNPs
A	A	AA	AG	cousin is universal match
A	G	AG	AG	both are universal matches
A	G	AG	GG	child is universal match, but the allele came from the paternal side
G	A	AG	AA	child is universal match, but the allele came from the paternal side
G	G	GG	AG	cousin is universal match
A	G	AG	GG	child is universal match, but the allele came from the paternal side
A	G	AG	AG	both are universal matches
A	G	AG	GG	child is universal match, but the allele came from the paternal side
A	A	AA	AA	the A allele is found in 85% of the population, so it's easy to match here
G	A	AG	GG	child is universal match, but the allele came from the maternal side
G	G	GG	AA	opposite homozygotes terminate the run of half-identical SNPs

Case D represents still another arrangement of alleles, with two rather short segments overlapping to create the appearance of one longer segment. FTDNA calls this a compound segment. Again, the child's match would not appear in either parent's list, so the segment as a whole is IBS. These short segments may be more frequent when both parents come from a population group with a limited number of founders, such as island populations, French Canadians, Mennonites, or Ashkenazi Jews. The DNA of the founders is recycled as various lines of descent continue to intermarry, although it is fragmented into smaller pieces as recombination occurs over many generations.

Case D – compound segment

maternal allele	paternal allele	child's genotype	putative cousin's genotype	
A	A	AA	GG	opposite homozygotes not part of the run of half-identical SNPs
A	A	AA	AG	cousin is universal match
A	G	AG	AG	both are universal matches
A	G	AG	GG	child is universal match, but the allele came from the paternal side
G	A	AG	AA	child is universal match, but the allele came from the paternal side
G	G	GG	AG	cousin is universal match
G	A	AG	GG	child is universal match, but the allele came from the maternal side
A	G	AG	AG	both are universal matches
G	A	AG	GG	child is universal match, but the allele came from the maternal side
A	A	AA	AA	the A allele is found in 85% of the population, so it's easy to match here
G	A	AG	GG	child is universal match, but the allele came from the maternal side
G	G	GG	AA	opposite homozygotes terminate the run of half-identical SNPs

Clearly, genotypes are a black box compared to haplotypes. However, a father/mother/child trio is not always available for phasing, and the raw data for the match is not easy to come by, especially in the world of genetic genealogy, where customers have an expectation of privacy.

Still, we may be able to make some indirect observations by collecting cases where one parent and a child have the same match. At Family Tree DNA, at least one long segment is required to declare a match,[7] but

segments as short as 1 cM are reported. A parent may not pass along every segment to the child (indeed, the chances are only 50-50), but every segment in the child should appear in the parent if it is IBD.

Figure 1 shows the results for a small convenience sample, with 269 segments of various lengths appearing in 19 cousins of six people. The relationships are unknown, but predicted to be 3rd cousins with a range of 2nd to 4th cousin. There is a clear break between the long segments used to declare a match and the large number of smaller segments. Each cousin had one (and only one) long segment that was found in the parent (a requirement to be included in this study),[8] but most of the short segments (even the two 5 cM segments) are not found in the parent and are likely to be IBS. Preliminary results for an even smaller dataset of *known* cousins showed multiple segments of intermediate sizes appearing in the gap, which would raise confidence in the prediction. Likewise, preliminary results for a smaller dataset of Ashkenazi connections appeared to show a different distribution of IBS segments.

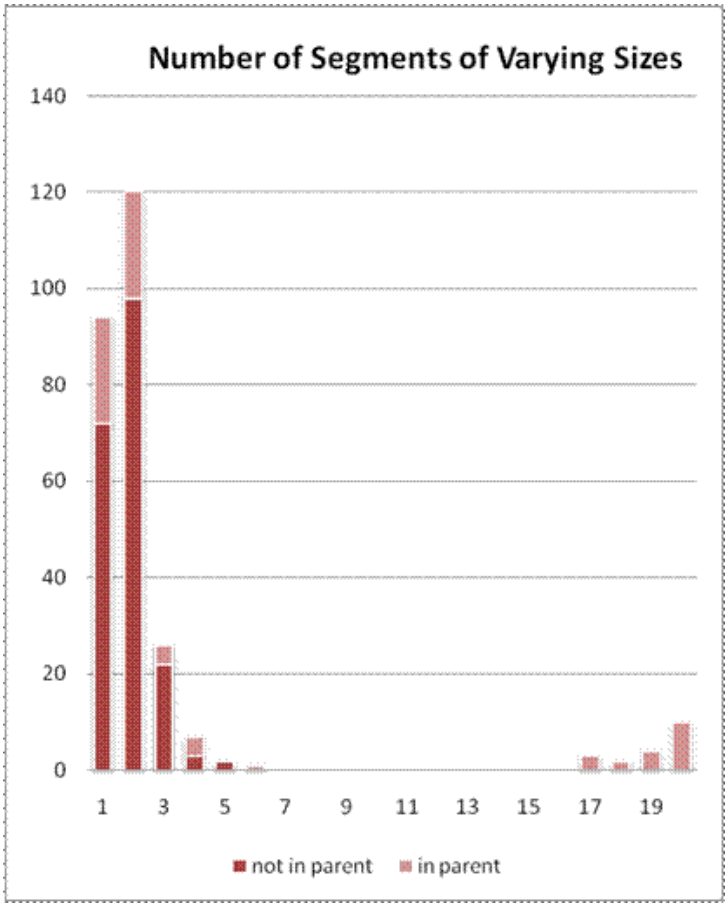

Figure 1 Distribution of segment sizes

Figure 2 displays the same data on a percentage basis.

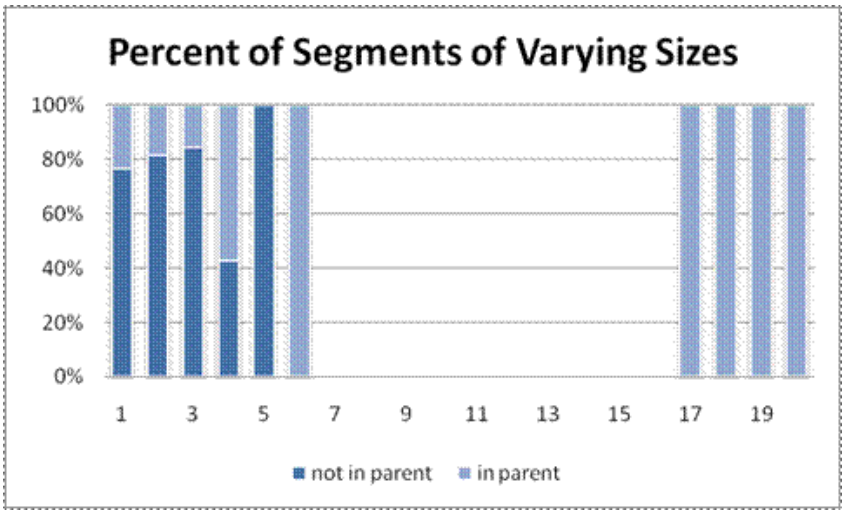

Figure 2

Learning how to handle small segments will be important for piecing together the contributions of our ancestors' DNA. More data may shed light on how much weight to give them in various scenarios. Some of the segments are gold nuggets waiting to be mined, but others will be fool's gold. See Appendix A for instructions on how to submit your data.

Appendix A

To add your data, log in to your FTDNA account for the child. Click on Matches, then on "Sort by Suggested Relationship." Find the names in the 3rd Cousin category (with a range of 2nd to 4th). This category is chosen because it is more likely to include a few segments of intermediate size, yet leave enough vacant space on the chromosome to expose IBS segments. It is also less onerous to complete the steps below for a limited number of matches.

Next, check the parent's account to see if the same name(s) show up. If so, go to Chromosome Browser and click on the match name(s). Download the list of segments to Excel. If there are more than five matches, you will need to create more files. I can consolidate those later.

Return to the child's account and repeat the process.

The Excel files will have the kit number in the file name, but you may change that if you wish. Do indicate which file has the child and which file has the parent. If you wish, you may anonymize the match names before sending them to me. Use the word child, mother, or father in the NAME column and create a code for the match name. Use that code for all segments in the child's file AND parent's file. Names will not appear in the summary data in any event. Please specify the relationship for each match name if known (which need not be exactly 3rd cousin), and also include any information about the parent's general ancestral background. If you have multiple combinations of parent/child duos, please submit just one set of data.

The Excel files are in CSV format and can be read with a text editor. If opened in Excel, the parent's table will look like this:

Table 1

NAME	MATCHNAME	CHROMOSOME	START LOCATION	END LOCATION	CENTI-MORGANS	MATCHING SNPS
mother	John Anon	1	97588657	99291699	1.71	700
mother	John Anon	1	224471080	226945145	1.16	600
mother	John Anon	2	234095657	235095463	2.65	600
mother	John Anon	3	45411945	48644652	1.18	600

mother	John Anon	5	11920393	40965300	32.58	6547
mother	John Anon	5	140787440	142379050	3.12	500
mother	John Anon	8	54314931	56987939	2.22	600
mother	John Anon	9	76706743	78024292	2.83	500
mother	John Anon	10	109828992	112176138	1.41	500
mother	John Anon	11	56530836	62144995	4.26	1400

The author will later add a column with calculated values for % OVERLAP to the child's table, the basis for creating the consolidated graphs in Figures 1 and 2.

NAME	MATCHNAME	CHROMOSOME	START LOCATION	END LOCATION	CENTI-MORGANS	MATCHING SNPS	% OVERLAP (added by author)
child	John Anon	3	15133155	17422414	2.34	600	0
child	John Anon	5	13450992	32489782	21.84	4147	100
child	John Anon	5	131679007	133142127	1.92	500	0
child	John Anon	5	140484616	142379050	3.37	600	84
child	John Anon	5	153466978	155311339	1.80	500	0
child	John Anon	7	95963251	98251592	2.80	500	0
child	John Anon	8	97296949	99224457	2.44	500	0
child	John Anon	11	58149527	60332717	1.86	500	100
child	John Anon	12	60030888	61695894	1.84	500	0
child	John Anon	17	44234078	46287827	3.47	600	0
child	John Anon	20	37377033	40001851	2.45	600	0

Table 2

Note that one of the child's segments is actually longer than the parent's – only 84% is present in the parent. Figures 1 and 2 include any segment with at least an 80% overlap. Rounding to the nearest block of SNPs (500 vs. 600 in this example) may exaggerate any discrepancies, but phased data (if available) will allow a closer approximation to the true length of the segment. Further analysis of this example, even with phased data available for just one party to the match, showed that the segment was no more than 495 SNPs in length. The missing markers were scattered in the zig-zag fashion illustrated in Case B.

The phased data for this pair also reveals that the 1.86 cM segment on chromosome 11 is IBS, even though both child and parent match John Anon. What appears to be one continuous run of SNPs when looking at the genotype breaks down into several dozen short fragments when using the haplotype (i.e., alleles the child inherited from the mother). ***Thus a segment that matches both the child and the parent based on genotypes is not conclusive evidence for Identity by Descent.*** The parent's genotype may be IBS with the match independently. In the absence of haplotype data, the presence of a segment in both parent and child is best used as a screening tool to identify the ones meriting further investigation.

The question may arise of whether the segments missing in the mother might be found in the father. It is a theoretical possibility, but it seems to be uncommon based on a few preliminary checks. Any two people chosen at random may have some short segments that appear to be IBD.
Extra Credit:

Another question is whether the threshold for declaring a match in the first place is high enough to guarantee IBD. An informative exercise for people who do have father/mother/child data is to count the number of matches found in the child but not in either parent.

An Excel spreadsheet can automate the process of comparing parent and child data. Open an empty spreadsheet and call it "Trio Matches." Label column A "Owner" and column B "Count." Download the child's

match list into a CSV file, open it as a separate spreadsheet, copy the data including the header, and paste it into column C1 of "Trio Matches." Type the word "child" in cell A2, press enter, then position the cursor in the lower right hand corner of cell A2 and drag down to fill in all the rows. Repeat with the father's match list (omitting the header), copying the data below the child's list in column C and filling in column A with the word "father." Repeat again with the mother's match list, filling in column A with the word "mother."

In cell B2, enter the formula

=COUNTIF(C:C,C2)

Press enter, then place the cursor in the lower right hand of cell B2 and drag down to fill in all the rows.

Sort the file by column A, then column B. Matches not found in either parent will have the number "1" in column B for the child. Copy the values for the longest block (column G) for these cases and send to me, along with the total number of matches found in the child (excluding the match with the parents, where the longest block will be very long). For instance, one child had 17/70 matches not found in either parent. The lengths of these were:

7.70
7.72
7.75
7.76
7.80
7.82
7.94
7.97
7.97
8.03
8.12
8.28
8.81
9.45
9.70
11.93
12.90

Ann Turner
DNACousins@aol.com

Disclosure

Ann Turner has a consulting agreement with the company 23andMe, Inc. The opinions expressed in this article are entirely her own.

Web Resources

23andMe Web Site
http://www.23andme.com

DeCodeMe Web Site
http://www.decodeme.com

[1] http://www.jogg.info/62/files/SatiableCuriosity.pdf

[2] http://www.familytreedna.com/landing/family-finder.aspx

[3] An online tool is available at http://www.math.mun.ca/~dapike/FF23utils/trio-phase.php

[4] The vast majority of SNPs are bi-allelic (two known alleles), but there are a few that have three or even four alleles.

[5] The absolute number of SNPs is not the primary consideration in determining whether a segment is IBD. Rather, it is a unit of length called a cM, which pertains to the probability of a segment being split up in the next generation. However, a minimum threshold of SNPs is required to allow ample opportunity to observe a contradiction

[6] It may also happen that the segment occurs in both the parent and the child, but the matching segment is longer in the child for this same reason.

[7] Empirical observations seem to show a minimum threshold of 7.7 cM and 500 SNPs.

[8] People who do have trio data may wish to tally the total number of matches appearing in the child but not in either parent. See the “Extra Credit” exercise in Appendix A.