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.

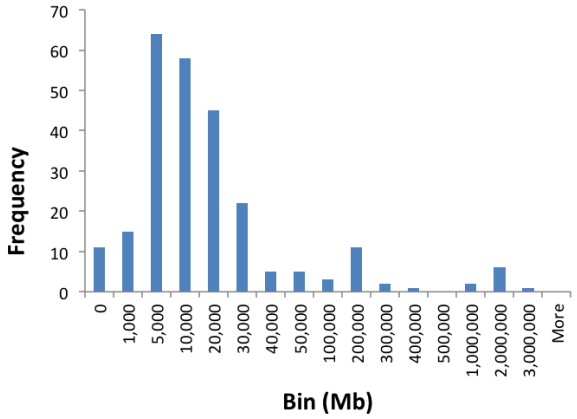


Figure 4. Distribution of 251 mismatch (gap) sizes in 86 parent/child comparisons at Gedmatch.com.

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

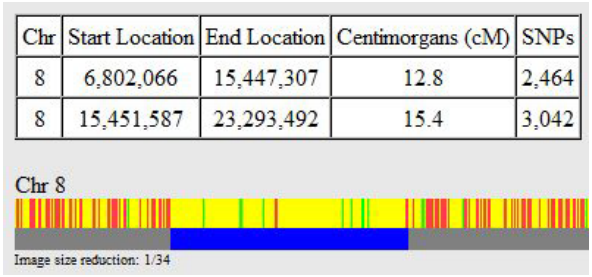
utilities. This is not to say they are false gaps – even mismatches on a single SNP could theoretically be due to a small deletion, although genotyping error rates could account for them as well. But the evidence for a microdeletion is much stronger when multiple SNPs are involved.

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.

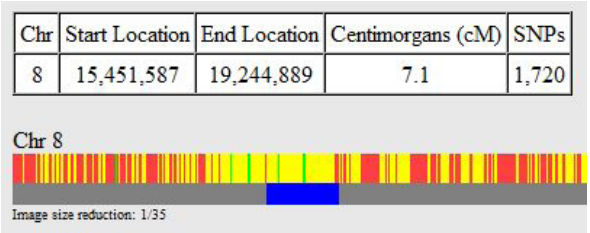
Figure 8 shows an example of a match found between a mother and a cousin, with two side-by-side segments separated by a small gap at the red bar. (The blue band appears continuous at this zoom level.) The daughter showed much the same segment boundaries, indicating that she inherited the deletion.



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.

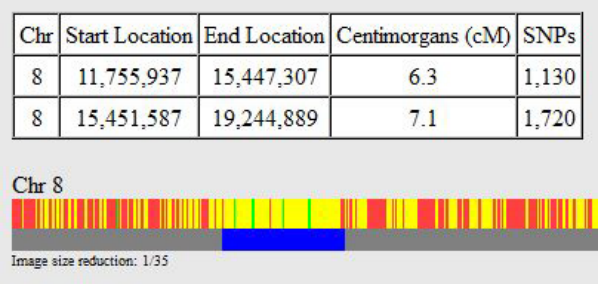
Figure 9 shows this region in a more distant cousin. Only the right-hand portion (starting at 15,451,587) registers at the default threshold of 7 cM, even though the amount of yellow and green in the left-hand side appears more prominent compared to the densely packed red bars outside of the segment.



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

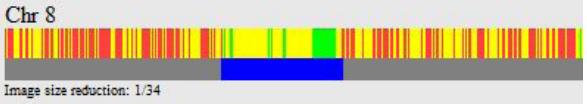


Figure 11. Match between the two cousins in Figures 8 and 9.

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