'SATIABLE CURIOSITY

Going Through a Phase: Haplotyping the Female X Chromosomes

'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.

Genetic genealogists have relied primarily on analysis of mitochondrial DNA for the straight maternal line and the Y chromosome (in males) for the straight paternal line. The inheritance pathway is unambiguous: we know the slots on a pedigree chart where our DNA results belong, even if we don't know the name of the person. These two forms of DNA are always in a "haploid" state, from the Greek word *haplous* meaning single.[1]

But the Y and mtDNA are only a small fraction of our DNA. What can we learn from other parts of our DNA, the X chromosome and the autosomes (non-sex chromosomes)? The autosomes come in matched pairs, one inherited from the father and the other from the mother, while the X chromosome is paired (diploid) in females and haploid in males. Our parents in turn passed on random portions of the DNA they inherited from their parents, so the inheritance pattern can zig-zag back and forth between males and females.

Genome-wide testing of many hundreds of thousands of markers is now available to the ordinary consumer from companies such as deCODEme[2] and 23andMe.[3] These markers are primarily SNPs (Single Nucleotide Polymorphisms, a substitution of one base A/C/G/T for another). But the analysis is complicated by the fact that any given stretch of our DNA could have come from any of a very large number of our ancestors.[4]

The difficulty is also compounded by that fact that we females can't even separate out which of our two X chromosome results came from our father's side of the family and which from our mother's side. Males have it lucky, since they know their single X chromosome came from the mother's side. The two alleles (alternative versions of a marker) comprising the female *genotype* are always listed in alphabetical order, as shown in Table 1, and we don't know which alleles are on the

Table 1
Reference SNP ID #, Chromosome (X), Base Position, Alleles

1	rs5990881	X	20446593	AG
2	rs4969758	X	20456331	GT
3	rs7876243	X	20473147	CT
4	rs1381266	X	20482613	AC

1 The term *haplotype* (the results from testing a set of markers located on a single chromosome) has actually been adopted from its original application. It was first used about 1969 in conjunction with the Human Leukocyte Antigen (HLA) system, a set of genes located close together on chromosome 6 and important for determining tissue compatibility for transplants. It was observed that if one member of a family had certain versions (alleles) for HLA-A and HLA-B, other members who matched the allele for HLA-A would almost invariably match the allele for HLA-B as well. This was evidence that the two alleles traveled together as a package on a single chromosome, whether inherited from the father or from the mother.

2 http://decodeme.com

3 http://23andme.com

4 For an animated illustration of the different pathways of inheritance, see http://www.smgf.org/pages/animations.jspx

same chromosome. This small set of SNP genotypes, located close together on my X chromosome, could represent a number of haplotypes (alleles located on the same chromosome).

For the first two SNPs, I could have inherited four different haplotypes: A-G, or A-T, or G-G, or G-T from my father, with the leftover alleles coming from my mother. Adding the alleles from the third SNP would double the possibilities: the C could go with any of the four haplotypes, and likewise the T, so now we're up to eight distinct haplotypes. The number doubles with each additional heterozygous SNP, making 2^4 or sixteen possibilities for a haplotype composed of just these four markers.

We know that the DNA from our more recent ancestors is passed down in rather long stretches, called haplotype blocks. Figure 1 is a descendant chart, showing that one of my X chromosomes (in red) will be composed of segments traceable to my paternal grandmother. Likewise, my male first cousin once removed will have inherited some segments from her. Some, but not all, of our segments will overlap.

Figure 2A shows graphically the actual overlap between my cousin and myself. Figure 2B and 2C compare me to two males of European ancestry, not known to be related to me. The narrowest green bands represent approximately one million bases where I am at least half-identical to them. I share some green bands with both unrelated males, although they occur at different positions on the X chromosome. These bands simply

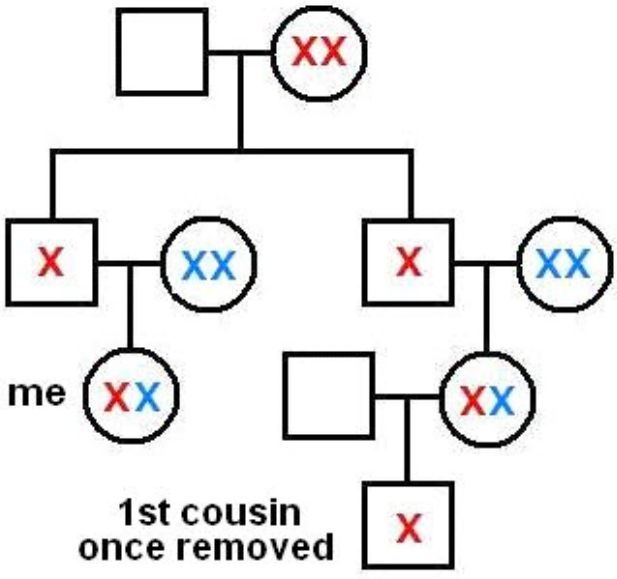

Figure 1. Descendancy chart for X chromosomes

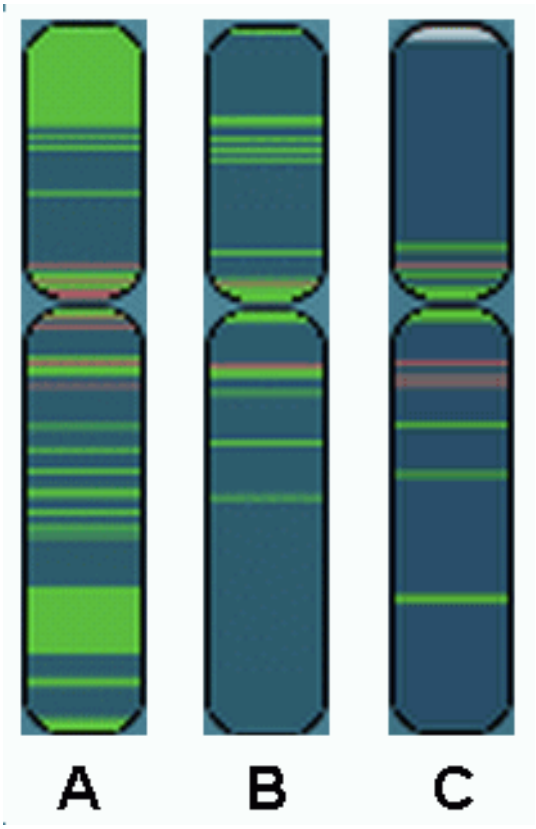

Figure 2. deCODEme "Compare Me" ideograms. Colored bars represent segments that are at least half-identical

reflect different parts of the European gene pool in general. The two broad green bands in 2A, one at the tip of the short arm and the other near the end of the long arm, are clearly more extensive than those of the randomly selected persons.[5]

By examining the raw data, I learned that the top green band in Figure 2A covered bases 214,201 to 21,962, 112. This region contains a run of 249 consecutive SNPs, 86 of which are heterozygous. The theoretical number of distinct haplotypes would thus be 2^{86} , which rounds off to 77,371,252,455,336,300,000,000,000. Yet I can look at my male cousin's haplotype and instantly convert my genotype into two haplotypes: one will match my cousin, and the other (which came from my mother's side of the family) will have the leftovers. The process of deducing which alleles come from the same chromosome is called *phasing,* often performed by

5 Parenthetically, the region around the centromere (the narrow waist of the chromosome) tends to show more similarity between people across the board. This section does not recombine as often.

software programs using a large number of population samples rather than the pedigree analysis developed here. Table 2 adds genotype results for samples in Figure 2A and 2B and divides my genotype into two phased haplotypes, P1 and P2.

Table 2
Genotype Data for 2A (cousin), Ann, Phased Haplotype 1 (Matching 2A from the Paternal Side), Phased Haplotype 2 (Deduced for the Maternal Side), and an Unrelated Male 2B.

	2A	Ann	P1	P2	2B
rs5990881	A	AG	A	G	G
rs4969758	T	GT	T	G	G
rs7876243	C	CT	C	T	T
rs1381266	C	AC	C	A	A

The male in Figure 2B has an overlapping green band, covering bases 18,359,428 to 21,638,884. His haplotype should match one or the other of my phased haplotypes, and in fact it does. If I did not have data from my cousin, I could still deduce my phased haplotypes by comparison with 2B, although I would not know which haplotype came from the paternal side and which from the maternal side. It would be exceedingly improbable to match that many consecutive SNPs by chance–we both must have inherited the haplotype block from a common ancestor. By comparing myself with a number of males, related or not, I could eventually phase a goodly part of my X chromosome.

Although this column uses diagrams from deCODEme to illustrate the process visually, there is enough overlap with markers used by 23andMe to make it feasible to merge raw data from the two companies. 23andMe's "Family Inheritance" feature is similar in principle to deCODEme's "Compare Me," but it does not highlight shared regions unless they are more extensive – about 10 million bases, enough to justify the "Family" aspect of the comparison. However, a lower threshold could be set for analyzing haplotype blocks in the raw data.

Male-to-male comparisons are also possible. Here the phase is already known, and the point of interest is whether they share extended regions of similarity, which would be evidence of descent from a common ancestor. Longer haplotype blocks would indicate more recent ancestry, while shorter haplotype blocks would be identical by descent from a more distant ancestor, perhaps even thousands of years ago. A collaborative project could perhaps develop a "dictionary" of haplotype blocks correlated with geographic information. A pilot study might pick a non-coding region of some optimum length and solicit data from people without raising concerns of revealing medically sensitive information. With the method described in this column, both males and females might be able to pool data, creating a larger sample size than used by many publications.

Ann Turner
DNACousins@aol.com