'SATIABLE CURIOSITY

SNPs on Chips:
A New Source of Data for Y-Chromosome Studies

'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.**

Single Nucleotide Polymorphisms (SNPs, pronounced snips) are mutations that substitute one base (A, C, G or T) for another, thus creating two possible versions of a DNA sequence, called ancestral and derived. The SNP mutation rate is so low that SNPs are sometimes called Unique Event Polymorphisms (UEPs), mutations that have occurred only once in all of human history. The U in UEP is an exaggeration, but the SNPs are rare enough that they can serve as landmarks for branching lines of descent from Y-Adam.[1] A series of SNPs can thus be used to define haplogroups and subhaplogroups, large clusters of males who share a common patrilineal ancestor thousands of years ago.

Until recently, tests for Y SNPs used a targeted approach, selecting a limited number of the best candidates for analysis. For instance, Family Tree DNA can often predict a haplogroup based on similarities in the STR haplotype (the most common test for genealogy, with marker names like DYS19). The SNP test is then used to confirm the prediction formally.[2] For certain haplogroups, follow-up tests are available.[3] Another test, the Marligen multiplex,[4] uses a two-stage approach. The first stage determines the major branch off the main trunk (thus not requiring an STR haplotype), and the second stage refines the haplogroup into subhaplogroups. Ethnoancestry[5] offers a smorgasbord of standard and experimental SNPs, which can be ordered singly or in any combination.

Now several companies[6] are offering scans for hundreds of thousands of SNPs, scattered across the entire genome. Tests for these SNPs are assembled on a platform commonly called a chip, because the original manufacturing process was similar to the one used to design computer chips. Although the emphasis is on the 22 pairs of autosomal (non-sex) chromosomes, the mass-produced genome chips also contain hundreds of Y SNPs, more than ever tested before in a simultaneous fashion.

At the time of this writing, a smattering of individuals have received results from two of the genome companies, 23andMe and deCODEme. The two companies have taken a different approach to SNP selection.

23andMe has added a number of custom SNPs to the off-the-shelf chip from Illumina. These were specifically chosen to cover much of the phylogenetic tree, in collaboration with Peter Underhill, discoverer of many of the SNPs. 23andMe reports on a total of 284 SNPs, with 230 of those registered at dbSNP, a centralized public

---

1 The current version of the ISOGG Y Phylogenetic Tree can be found at http://www.isogg.org/tree.

2 Ordering information is available only on the personal results page of people who have obtained STR results.

3 http://www.familytreedna.com/deepclade.html

4 DNA Heritage is one company offering this test Background information can be found at: http://www.dnaheritage.com/ysnp.asp

5 http://www.ethnoancestry.com/custom.htm

6 http://www.23andme.com, http://www.decodeme.com, http://www.seqwright.com, http://www.geneessence.com

7 http://www.ncbi.nlm.nih.gov/projects/SNP/

repository at the National Center for Biotechnology Information (NCBI).[7] Those SNPs are assigned an rs number (Reference SNP), and the records may contain additional information, such as their frequency in various populations. The remaining 54 SNPs are assigned an "i" number (internal to 23andMe).[8] With this suite of SNPs, 23andMe is able to assign very derived haplogroup labels, such as R1b1c9a (2007 nomenclature).[9]

In contrast, deCODEme does not assign such derived haplogroups, yet it includes more Y SNPs in the raw data, a total of 858. These SNPs are all in dbSNP and come from a large variety of sources. They have simply been observed to occur in some setting or another, and they are not necessarily vetted for their placement in a tree structure.

That immediately makes the curious genetic genealogist ask the question "CAN these SNPs be placed on the phylogenetic tree? In fact, would the SNPs revise the tree, uniting some branches or adding some twigs at the tips?" The genome tests are expensive (in the $1000 range), but if a few pioneering individuals share their raw data, all may benefit from the insights gained by comparing even a few people.

In fact, this cooperative endeavor is taking place informally right now. Since R1b1c is very common, the first interesting discovery involved a SNP in that haplogroup, rs34276300, which was found to be ancestral in one branch of R1b1c and derived in several other branches. Two companies, EthnoAncestry and Family Tree DNA, began offering a test shortly thereafter. Thus the methodology has gone full circle, from targeted tests to genome scans and back to targeted tests!

A more systematic effort to collect genotype data, from a broader variety of haplogroups, would undoubtedly speed up the process. I am proposing a centralized spreadsheet of raw data, which can serve as an "open source" for data miners.[10] There are many creative ways to analyze and present data, and the spreadsheet will include links to websites that have made use of the resource. In addition, the spreadsheet includes a master table with a consolidated list of all the SNPs tested by the various companies, together with the YCC 2008 nomenclature for haplogroups associated with each SNP (Karafet, 2008). [11]

Contributors to this collaborative effort may write to me for further instructions on extracting their genotypes from the complete genome scan. No medical implications of the Y-SNP portion are known at the present time, although fertility problems might be evidenced by several consecutive no-calls in the region of genes responsible for sperm production.[12] Contributors may remain anonymous, but if they are willing to be contacted for further information (such as STR haplotypes, results of previous SNP tests, ancestral origins, or other points of interest), the analysis will be all the richer.

Although 23andMe customers may feel they already have all the answers, there are other points of interest, and input is solicited from them as well. For instance, there is the question of the actual uniqueness in UEP– how often do parallel and reverse mutations occur? The targeted SNP tests skip over large numbers of SNPs, and perhaps one will show up when more "irrelevant" markers are included. Preliminary results have also revealed a curious phenomenon: one marker has been heterozygous (exhibiting two alleles) in several R1b1c individuals, but homozygous (one allele) in Haplogroup I. The reason for this is not known.

The sheer quantity of raw data is unprecedented, and genetic genealogists are in a position to help interpret it.

Ann Turner
DNACousins@aol.com

Web Resources

http://dnacousins.com/SNPs_on_Chips.xls
http://dnacousins.com/SNPs_on_Chips.zip
Y-SNPs on Chips Database

References

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase the resolution of the human Y chromosomal haplogroup tree. *Genome Res*, 18:830-838.

King TE, Bosch E, Adams SM, Parkin EJ, Rosser ZH, Jobling MA (2005) Inadvertent diagnosis of male infertility through genealogical DNA testing. *J Med Genet*, 42:366-368.

Update (October 2008): 23andMe began including a total of 1358 Y SNPs in raw data downloads by May, 2008. Version 2 of their chip was released October 2008 and contains 2042 Y SNPs.

8 Gareth Henson has mapped those SNPs to the M numbers, as they appear on the ISOGG (2007) Y tree. Those data are included in the spreadsheet discussed in the text.

9 Occasionally, some SNPs cannot be "called," or assigned a value, and the assignment may not be quite as deep, e.g., R1b1c*.

10 See Web Resources.

11 Not all markers in this tree have rs numbers.

12 A large-scale deletion, approximately 4 million bases roughly between positions 23,000,000 and 27,000,000, was described by King et al. (2005). This region contains a commonly tested Y-STR marker, DYS464. The estimated frequency in the general population is about 1 in 4000.