'SATIABLE CURIOSITY

Up Hill and Down Dale in the Genomic Landscape: The Odd Distribution of Matching Segments

'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.

Ann Turner
DNACousins@aol.com

More genetic genealogy companies are offering genome-wide tests. The latest entrant is Family Tree DNA's Family Finder,[1] 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.

It is both a forte and a flaw of the human mind to look for patterns – oddities that stand out against a noisy background. Sometimes we actually create patterns out of random events, but sometimes our observations reveal an underlying structure. A number of people have reported that their matching segments aren't evenly distributed over all the chromosomes. Instead, they see an abundance of segments on some chromosomes and a paucity on other chromosomes. Are these just "segments of our imagination," based on a small amount of data, analogous to seeing seven heads when tossing a coin ten times and claiming that the coin absolutely must be biased?[2] Or are they clues to previously undetected features of the genomic landscape?

No one individual has enough data to settle this question, but collaboration opens up more possibilities. Details about the sizes and lengths of matching segments can be downloaded into a spreadsheet (see Appendix A), and results for one individual can be merged into a master file.

I have prepared a spreadsheet (Supplement 1) to collect data of this type. Columns A to G (as shown in Table 1) are the same as the file downloaded from FTDNA (with actual names replaced by a code). The remainder of the spreadsheet performs calculations on the raw data.

One obvious, perhaps trivial, hypothesis is that longer chromosomes will show more segments. I am taking this to be a given, and calculating the "expected" number of segments for a chromosome to be proportional to the chromosome length in cM.[3] Table 2 shows the results for two datasets combined.

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

[2] This distribution is "expected" to occur about 11.7% of the time with a fair coin.

[3] The cM (centiMorgan) unit is a measure of how often a region splits up by recombination during the creation of eggs or sperm. The values in Table 2 are taken from Rutgers' "second generation" map at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2099587/. The measure is empirically derived, and numbers may vary somewhat in different study populations.

Name Code	Match Name Code	Chr	Start	End	cM	# SNPS
P001	P001-M001	4	111,444,399	115,754,265	5.11	906
P001	P001-M001	6	21,638,348	23,428,513	2.85	604
P001	P001-M001	8	82,949,582	86,998,001	1.72	501

Table 1

Chr	cM	observed	expected	obs/exp
1	286	159	131.3	1.21
2	265	166	121.4	1.37
3	224	88	102.7	0.86
4	215	125	98.5	1.27
5	209	77	95.7	0.80
6	196	200	90.0	2.22
7	188	66	86.3	0.76
8	169	126	77.4	1.63
9	167	106	76.7	1.38
10	175	73	80.3	0.91
11	162	98	74.2	1.32
12	176	94	80.6	1.17
13	132	37	60.5	0.61
14	125	57	57.5	0.99
15	132	30	60.5	0.50
16	133	28	61.2	0.46
17	139	26	63.7	0.41
18	130	35	59.5	0.59
19	111	4	51.0	0.08
20	114	36	52.5	0.69
21	69	9	31.7	0.28
22	80	10	36.8	0.27

Table 2

With this a priori assumption, the ratio of observed to expected mutations (obs/exp) would be exactly one for every chromosome. What happens when we graph the ratio? Instead of a straight line across the graph at Y = 1, we see a roller coaster.

Figures 1 and 2 show the results for two individuals. P001 is higher on chromosome 1 than chromosome 2, and P002 reverses the pattern. Figure 3 shows the data merged into one file. The merged data begins to flatten the curve for some chromosomes, but some peaks still stand out, notably the one on chromosome 6.

Chromosome 6 houses a set of genes known collectively as the Major Histocompatibility Complex., important for the immune system's role of distinguishing the self from a foreign invader. The genes, including the HLA (Human Leukocyte Antigen) genes tested for organ transplant compatibility, are densely packed into a region of about 3,600,000 bases (3.6 Mb) at positions 29,750,000 to 33,100,000. They harbor an enormous amount of variability, yet paradoxically, the recombination rate is low. Consequently, variants in different parts of the MHC often travel together as a package, a haplotype.* As a

consequence, a "conserved extended haplotype" may occur in large numbers of people. A recent study showed that the single most common haplotype was found in 8.7% of the subjects, and the ten most common haplotypes accounted for 30% .[4] From a genetic genealogist's perspective, similarities in this region may be too general to have utility in identifying recent relationships. However, if the number of matches in the region is on the low side compared to most people, the haplotype may be more distinctive and thus more informative.

Because of these biological underpinnings, it is quite possible that the peak on chromosome 6 will survive even after merging larger numbers of records. Will other peaks also stand out, or will they be eroded into flat valleys along the Y = 1 line? There are other candidates. Although the term "conserved extended haplotype" seems to be associated specifically with the MHC, the concept could apply to other regions. For instance, an inversion around the centromere of chromosome 9 may have reduced variability, since it does not align well with the more common version during the recombination process.[5]

And how about the original hypothesis, that the number of segments is proportional to the length of the chromosome? There seems to be a trend for matches in the smaller chromosomes to be underwater of the Y = 1 line. Is this just an accidental property of the first two datasets? More data will clarify the picture.[6]

More data – more data – more data. That is the mantra, and it is essential for developing a population baseline for comparison. But it leaves an individual in limbo, wondering if his personal peaks and valleys have any significance. It would be difficult to acquire the multi-generational data required, but a glance at a grandchild / paternal grandfather comparison (Figure 4) shows that the matching segments aren't distributed equitably over all the chromosomes to begin with. Some chromosomes match the grandfather across their entire length, while other chromosomes miss the grandfather's contribution entirely (and must come from the paternal grandmother). For instance, chromosome 2 in the granddaughter is every bit as good as her grandfather's: if her grandfather has a match on that chromosome, so will she. That effectively pushes her two generations closer to the common ancestor, and quadruples the chance of locating a cousin for that chromosome compared to the "average" behavior, with a 50% chance of losing a segment with each generation. Perhaps some of the peaks and valleys are an inevitable consequence of the lumpy distribution of DNA segments in the most recent generations.

I will collect contributions of data over the next several months and publish the results in the next edition of JoGG. Please see Appendix A for more details on how to prepare the data for submission.

---

[4] Szilágyi A, et al; Frequent occurrence of conserved extended haplotypes (CEHs) in two Caucasian populations. Mol Immunol. 2010 Jun;47(10):1899-904.

[5] http://en.wikipedia.org/wiki/Chromosomal_inversion

[6] Graphing data for the whole chromosome is obviously a crude measure, intended to illustrate the general method. The spreadsheet has a function for highlighting records that overlap some or all of a shorter segment., as chosen by the user. Note that some segments may appear to be exactly identical in size and position, but that is an artifact of the computation procedure and has no special significance beyond the fact that overlap is present.

* A person's results are reported as a genotype (two alleles for each marker, e.g. AG). It is not possible to separate the genotype into its two haplotypes without data from the father and the mother. After submission of this column, I had the opportunity to analyze haplotype data obtained from some father/mother/child trios. In some cases, a matching segment between the child and a cousin disappeared. It was actually cobbled together with small bits and pieces of contributions from the father and the mother, which accidentally combined to create the appearance of a longer segment. The shorter segments in the spreadsheet may thus be artifacts, but I will still compile the data and present results for shorter and longer segments separately.

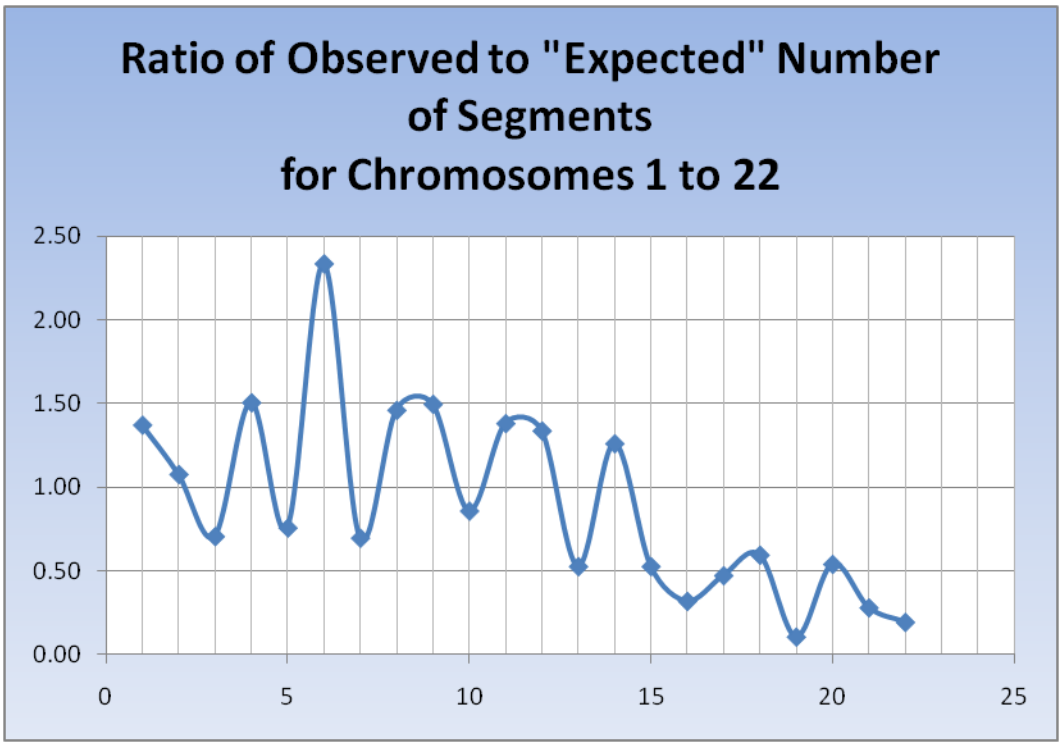

Figure 1: P001 in supplemental data

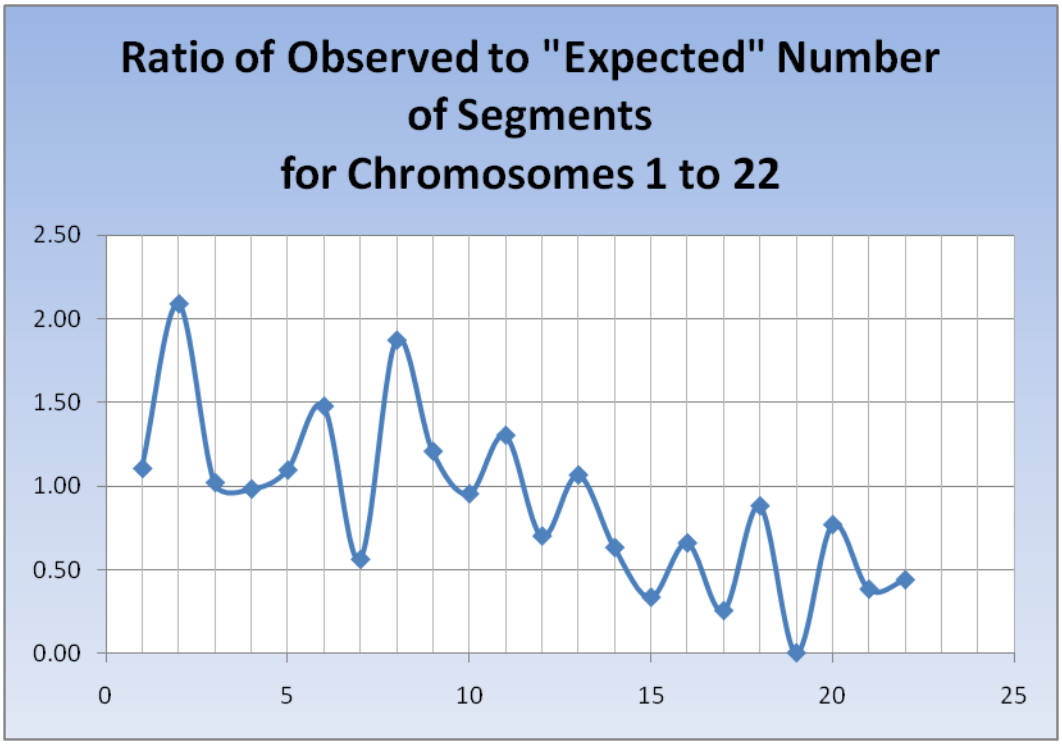

Figure 2: P002 in supplemental data

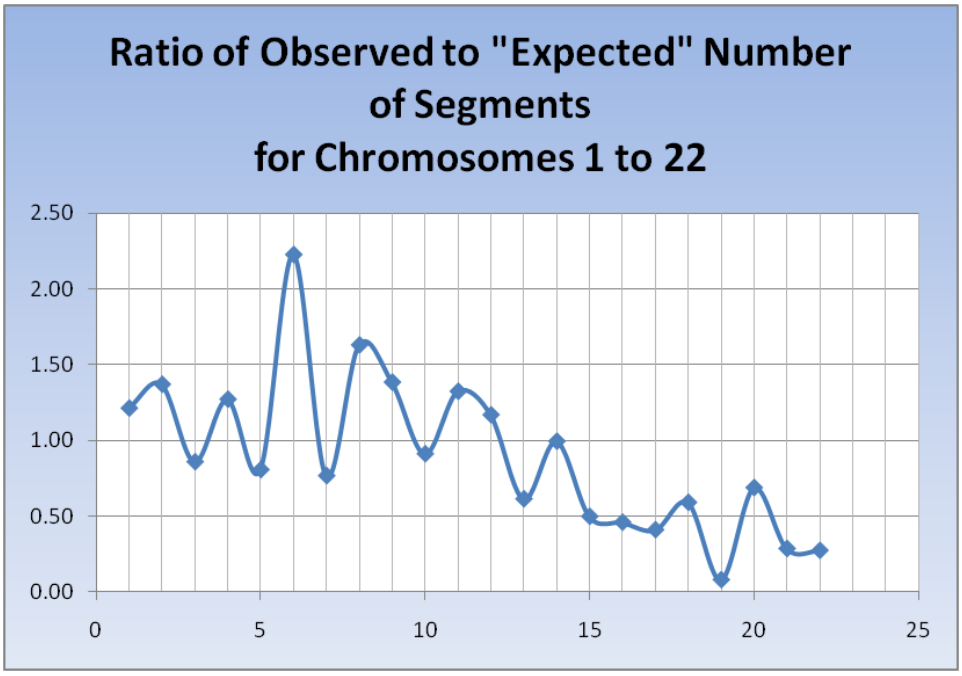

Figure 3: P001 and P002 combined

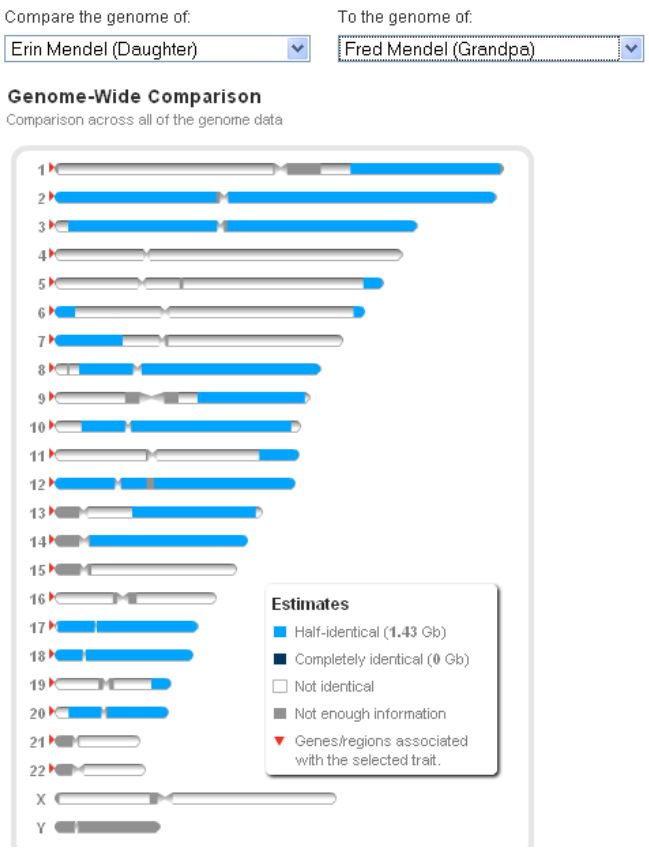

Figure 4: Family Inheritance Diagram, http://23andMe.com (visible with a demo account)

Appendix A

Download http://dnacousins.com/Segments.zip, open the “Segments” file, and switch to the "your data" worksheet. The “your data (2)” worksheet is a spare copy in case you need to start fresh.

Log into your FTDNA account.

Select Chromosome Browser in the Family Finder section.

Select up to three people.

Click on the "Download to Excel" link in the chromosome diagram.

Allow your browser to open the file, or save it to your hard drive and open it yourself.

Copy just the data rows (not the header) to the clipboard. (Position your mouse in cell A2, drag it to the lower right-hand corner of the data, and type Ctrl-C.)

Position your cursor in cell A2 (or first unused row) of the Segments.xlsx file.

Paste the data (Ctrl-V) into Segments.xlsx

Deselect the three people and highlight three more people. Repeat the above instructions until you have downloaded ALL of your data. (There is no provision at present for downloading all of your data as a batch.) Omit any relatives who are first cousins or closer. Please don’t “cherry-pick” just those matches where you’ve already identified multiple segments. If it is too tedious to complete the whole download, just pick a few of the pages and download everyone on them.

Excel will automatically perform the calculations and diagram the results as you go.

If you do not have Excel 2007, the spreadsheet will open in older versions. The format will be slightly different, and you will not be able to use a drop-down list in the column headings to filter your results, but the calculations will be the same.

You may enter any segment position in columns M3 to O3 to calculate the amount of overlap found among your matches. The default values are for the MHC on chromosome 6.

To submit your data, make a copy of your spreadsheet to save for yourself, then replace the names of your matches with a code (for privacy reasons).

Use M001 for the first person, M002 for the second person, etc. You can use the "Replace" function if you wish, but it's almost as fast to enter the code once, then copy and paste it into the rows below.

Leave your name in the spreadsheet. I will replace it with a code before publishing the merged data, but I may need to contact you with any questions.

E-mail to me, DNACousins@aol.com.

I will upload the most recent version of merged data to the JoGG website on a periodic basis, depending on the amount of data submitted. The file will be date-stamped as Segments_2010_08_17.xlsx, etc.