Single Segment Recombination Simulations

James Sims

Date: February 27, 2017

Summary

Genetic genealogists need a very basic appreciation of the statistics of random events in order to effectively use DNA segments with low cM values for DNA matching, especially when there is only one shared DNA segment. This document provides a description of random events in the context of recombination events affecting autosomal DNA as it routinely undergoes this process as part of the reproduction. The concepts of centiMorgan (cM), outcome distributions and segment survival in a genealogical time context are discussed.

Introduction

Most people, including genetic genealogists, have not studied statistics in some formal way. In spite of this, most people are able to get through life and conduct genealogy projects effectively. DNA analysis for ancestry in a direct-to-consumer sense is a 21st century phenomenon, one that carries with it some relatively unfamiliar mathematics including the statistics of random events. When genealogists add DNA analaysis to the list of information sources they use in support of their genealogy projects, they can wade into unfamiliar territory and make errors. For the most part, direct-to-consumer DNA testing companies such as 23AndMe, AncestryDNA and Family Tree DNA do most of the heavy statistical and mathematical lifting for their customers with regard to setting autosomal DNA matching thresholds. With the advent of online tools such as GEDmatch, it is possible for genetic genealogists to tinker with matching thresholds in the quest for more matches, and for the purpose of hypothesis testing. This document was written for those people who want to change matching thresholds, in particular, lowering the cM matching threshold.

Human autosomal DNA undergoes recombination in every generation as part of the reproductive process. It is this recombination that accounts for the findings that immediate family members share more DNA than they do with their cousins, and close cousins share more DNA than do distant cousins. Immediate family and close cousins tend to share DNA segments with large cM values. As the generations pass, ever more distant cousins share less and less DNA. Not only is the quantity of shared DNA less, the physical size of the shared DNA segments become smaller and smaller as recombination events in every generation break up what were longer segments of shared DNA. Very distant cousins will share, if they share DNA at all, segments with low cM values.

Genetic genealogists who pay attention to cM values of shared DNA segments tend to talk about cM values as if they are measures of physical length of DNA. This is incorrect. cM values refer to the recombination probabilities and not to physical length. The unit of physical length genetic genealogists should use is base pairs when they talk about length. It was not until about the year 2000, that scientists were first able to compare the physical map of chromosomes of complex organisms measured in base pairs with the recombination maps measured in cMs. They found that there was no fixed relationship between the length of the DNA segment measured in base pairs and the cM value of the segment. Different fruit fly chromosomes, for example, had different numbers of base pairs per cM. The two arms of each fruit fly chromosome differed in the number of base pairs per cM. This is also true in humans.

Because there is no one-to-one relationship between cM value and physical length measured in base pairs, genetic genealogists tend to focus on cM values of DNA segments rather than physical length. Recombination probabilities measured in cM are useful to genetic genealogists and are the subject of this document. A DNA segment with a cM value of 1 has a probability of 0.01, or a 1 percent chance of undergoing an odd number

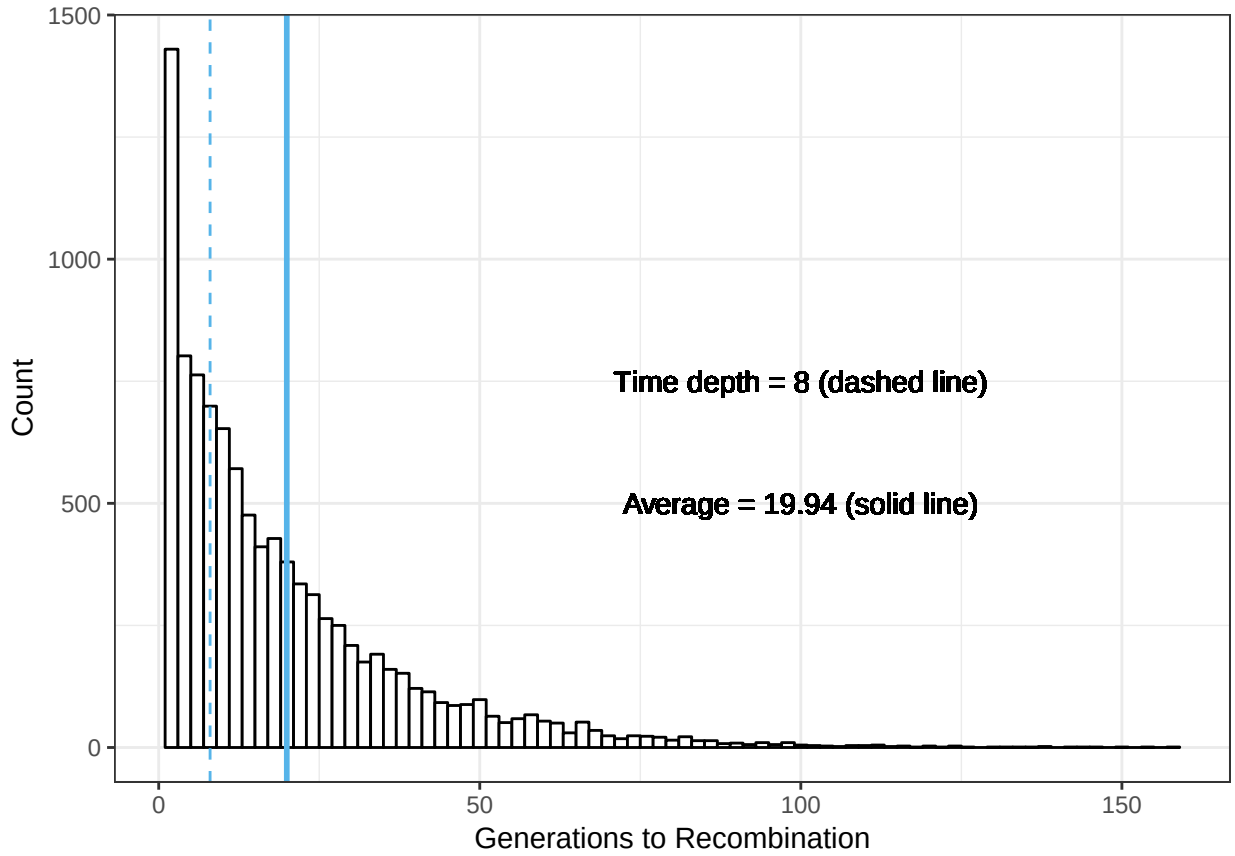

Figure 1: Histogram showing results of 10,000 simulated experiments of a 5 cM segment of DNA undergoing recombination. Solid blue line shows the average number of generations to recombination. The dashed blue line shows an arbitrary genealogical time depth of 8 generations.

of recombination events per generation. Likewise, for example, a DNA segment with a cM value of 5 has a 5 percent chance of undergoing an odd number of recombination events per generation.

Simulating recombination events

Consider the following thought experiment as one way to simulate recombination events for a DNA segment with a cM value of 5. Let a box containing 5 green marbles and 95 white marbles represent all possible outcomes of an experiment. By definition, there is a 5 percent chance (5/100) of drawing a green marble from the box with a single draw. Let the act of drawing one marble from the box represent the passage of one generation. To keep the probabilities the same for each draw (in each generation), the drawn marble is returned to the box before the next draw. To keep the experiment fair, we will draw marbles while blindfolded, and we will mix the box of marbles thoroughly between draws so the draws will be a random marble from the box. As we draw marbles (as the generations pass), we keep track of how many draws it takes to draw one green marble. Drawing a green marble means a recombination event took place, and the DNA segment broke up as a recognizable unit of inheritance due to recombination. At the point of drawing a green marble, the experiment ends. At the end of the experiment, we add up all the draws it took before a green marble was drawn. This experiment needs to be repeated a large number of times before a clear picture will emerge about how the recombination events are distributed over the generations.

This type of experiment is relatively easy to code on the computer as a simulation. The following discussion applies to using computer simulations to understand recombination events in a genealogical context. Figure 1

shows the distribution of results when you ask the question how long does it takes for a DNA segment with a cM value of 5 to undergo recombination measured in generations. In this figure, the bin width was set to 2. This means that each bar on the chart contains the number of experiments (count) for two generations. For example, the first bar on the left side of the figure shows how many times it took only one or two generations for the 5 cM DNA segment to undergo recombination. The next bar to the right shows how many experiments required only 3 or 4 generations before the recombination event occurred. Note the wide range of outcomes, including a few of more than 150 generations! In some experiments it took just a few generations for the segment to undergo recombination, but in many experiments, it took many more generations. The calculated average for the outcome distribution was 19.9, which is very close to the theoretical value of 20.0. This gives us some confidence that the simulation is not flawed mathematically.

When genetic genealogists look for matches with people who "share" only one DNA segment, the best of all outcomes is to find a single shared segment with a low probability of survival within the genealogical time frame of interest. Finding a shared DNA segment with a low probability of survival means that segment is more likely to have arisen within the genealogical time frame of interest compared to segments with higher probabilities of survival, which may be much older. In Figure 1, the 5 cM segment has a 66.59 percent survival at a time depth of 8 generations, meaning most of the recombination distribution is outside the genealogical time depth of interest. The problem genetic genealogist have today trying to use a 5 cM segment as a match at this time depth is they can't tell by looking at *a particular segment* if they are looking at a recent event or a very old one (where in the outcome distribution the observation belongs). One way to deal with this is probabilities. There is a 66.59 percent chance that the 5 cM segment of interest is not within the genealogical time frame of interest in this example. This being the case, using a segment like this to call a match a match is likely to be wrong 66.59 percent of the time. Not good!

It should be emphasized the foregoing example and figure applies only to situations in which there is only one "shared" DNA segment. That's the focus of this document—trying to decide if a single "shared"" segment constitutes good evidence of a match due to a shared ancestor within the genealogical time depth.

Simulated experiments just like the one that generated Figure 1 can be conducted easily on the computer in which segments vary in cM value and the genealogical time depth of interest is varied. Figure 2 shows the results of 10,000 simulations of segment recombination for 23 different cM values. The resulting segment recombination distribution for each cM value was then "sliced" at six different genealogical time depths from 8 to 50. Each curve (lines connecting data points) in Figure 2 is a survival curve for one of the genealogical time depths. If a genetic genealogist wishes to call a match with a distant cousin within the last 15 generations, a single shared segement of about 18 cM could be used to argue in favor of a called match saying making the match call will be correct about 95 percent of the time. However, if the genetic genealogist wants to argue for a match within 10 generations, finding a single shared segement of about 26 cM is needed to be correct about 95 percent of the time.

If the genetic genealogist is willing to live with more uncertainty in calling a match, smaller segments might be used with caution. For example, at the 10 percent survival level (yellow line in Figure 2), one might cite a 15 cM segement as supporting a single segment match at the 15 generation time depth or a 20 cM single segement at the 10 generation time depth. In these situations, a genetic genealogist might argue for being correct in calling a match a match about 90 percent of the time.

The adjectives the author might use when describing matches base on Figure 2 are: (1) **strong** evidence for single segment matches called at or below the green line, (2) **moderately strong** evidence for single segment matches called at or below the yellow line, but above the green line; (3) **weak** evidence for single segment matches called at or below the orange line, but above the yellow line; (4) **speculative** matches called at or below the red line, but above the orange line; (5) **no** evidence for single segment matches above the red line.

When citing evidence of single segment matching, at a minimum, genetic genealogists should cite the evidence: the match is based on a single segment; state the cM value of the segment; state which chromosome the match is on and the chromosome start and stop locations in base pairs and the number of shared SNPs. If you want to categorize the strength of the evidence as strong, moderately strong or weak, for example, do so but also include the numerical scale and how it was derived, or more simply cite the source of the scale. The following hypothetical paragraph might be written by an author arguing a genealogical point and

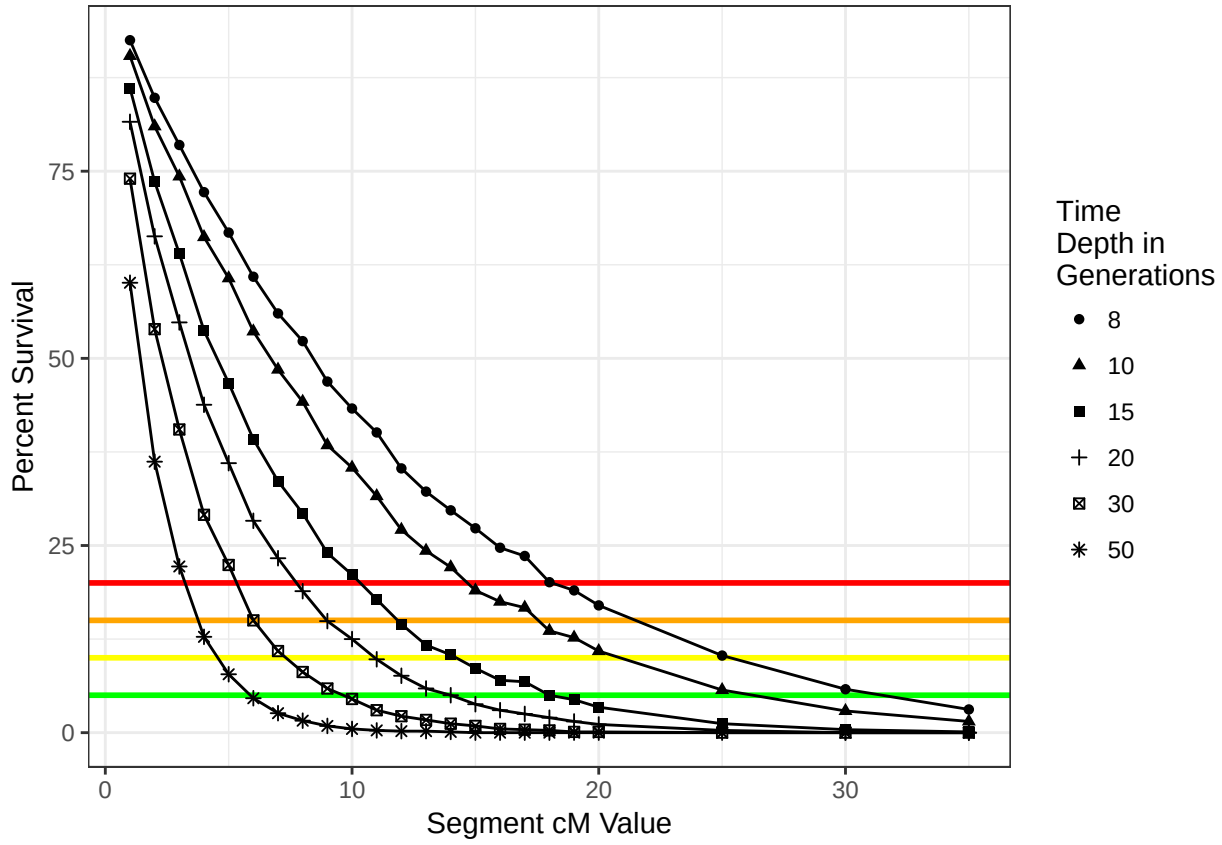

Figure 2: Survival curves for DNA segements of varying cM values at six genealogical time depths. Selected survivals are shown by green line, 5 percent; yellow line, 10 percent; orange line 15 percent; red line 20 percent.

referencing this document as the source of the evidence scale. By describing the evidence first, and then the interpretation in the following manner, if it turns out the interpretation is not quite correct, readers in the future will still have the primary evidence, and can use it in the future when the understanding of DNA evidence is more sophisticated.

Jane Smith Doe is a well documented descendant of Capt. Imhur Ancestor[1]*. The captain is her 7th great grandfather. Jane descends from the captain's oldest child Imhur Ancestor, Jr. I have a strong paper trail back to a woman named Ima Ancestor*[2]*. After a reasonably exhaustive search of the paper records, I am unable to prove by paper records that Ima Ancestor is the daughter of Capt. Imhur Ancestor. However, I can make a strong circumstantial case based on paper records that Ima Ancestor may be the daughter of Capt. Imhur Ancestor. If true, Capt. Imhur Ancestor was my paternal 6th great grandfather. According to autosomal DNA matching at Family Tree DNA, Jane Smith Doe appears as a paternal match for me based on the phasing of my DNA data with two paternal first cousins. Jane Smith Doe and I and one of my first cousins have a triangulated DNA match on chromosome 16 starting at position 12619984 and ending at position 30071334. According to Family Tree DNA, the cM value for this DNA segment is 25.24 and the number of shared SNPs is 3970. This segment on chromosome 16 is the only segment of DNA my paternal first cousin and I share with Jane Smith Doe. On the basis of single segment survival curves generated by computer simulations, the single segment match I have with Jane Smith Doe constitutes moderately strong evidence that Jane Smith Doe and I share a recent common ancestor within the last eight generations*[3]*. After a review of my genealogy and that of Jane Smith Doe, there appears to be no other families in common other than the Ancestor family back to Ima Ancestor for me and back to Capt. Imhur Ancestor for Jane Smith Doe. The DNA evidence in this case is moderately strong, and supports the strong circumstantial case for Ima Ancestor being the daughter of Capt. Imhur Ancestor.*

Methods

This pdf document was authored in its entirety using the free version of RStudio, Rmarkdown and the open-source statistical programming environment for data science known as R. All of the computer code used to generate this document including the figures is in the recombination_sim.Rmd file posted online at the author's GitHub account[4].

Hardware & software	Version	Use
Apple, Inc. iMac computer	Mid 2015	hardware
OSX	10.11.6	operating system
RStudio[5]	1.0.136	authoring Rmarkdown & R programming
R[6]	3.3.2	R programming environment
dplyr package[7]	0.5.0	dataframe manipulation in R
ggplot2[8]	2.2.0	plot charts
knitr package[9]	1.15.1	authoring Rmarkdown
MacTEX[10]	20161009	render Rmarkdown to pdf

[1]Supplemental application of Jane Smith Doe (1111111, Add Volume 3000) on Imhur Ancestor (A222222)," verified October 5, 2003; National Society Daughters of the American Revolution, Office of the Registrar General, Washington, D.C.

[2]Membership application of Mary Jane Smith (3333333) on Ima Ancestor (A3333333), approved April 19, 2005; National Society Daughters of the American Revolution, Office of the Registrar General, Washington, D.C.

[3]Sims, James, *Single Segment Recombination Simulations*, published online, February 27, 2017, accessed February 27, 2017, https://github.com/simsj/recombination

[4]Sims, James, *Single Segment Recombination Simulations*, published online, February 27, 2017, accessed February 27, 2017, https://github.com/simsj/recombination

[5]RStudio web site

[6]R-project.org web site

[7]dplyr package on CRAN

[8]ggplot2 package on CRAN

[9]knitr package on CRAN

[10]MacTEX web site