Statistical Methods 201-DDD-05 Final Project Report

Part 2: Scientific Essay on Modelling the Most Recent Common Ancestor

# Exhaustively Understanding the Most Recent Common Ancestors and Estimating the Time to the Most Recent Common Ancestors using Recently Proposed Computational Models

Nguyen Hoang Anh

John Abbott College

Montréal, CANADA.

## 1. Introduction();

On the internet, there is a site called the "Mathematics Genealogy Project" which can be used to draw an ancestral tree for mathematicians who have published papers. It works by looking into a vast database of papers and profiles to see who studied under whom, then we can manually work our way back to the earliest mathematicians. For example, our instructor for this course is Dr. Luiz K. Takei who got his Ph.D. at McGill University (Montréal). His advisors are Henri Rene Darmon and Eyal Z. Goren, looking at the advisors of the advisors of those two advisors, etc. You get the idea. We finally arrive at a rather peculiar mathematician, a German person named Carl Friedrich Gauß. Thanks to this database, we now know that our instructor is a descendant disciple of this German guy. Even though saying we (students in this course) are also descendant disciples of Gauss is a far-fetched idea in and of itself, it is interesting to see the long winding road down the mathematical timeline leads to a bunch of cégep students in Montréal.

Our fascination about the intrinsic connection between the past and the present does not initially stem from the curiosity of finding out who is a student of whom, but rather, collectively as the human race, where did we all begin? Is there a cut off between present-day us and the first *homo sapiens* appeared on this Earth? We raise the question of whether or not this Vietnamese international student is related to those Caucasian Canadian students sitting around him. While this question might raise eyebrows in front of leftist maxima about the alleged racial undertone ("we don't see colors"), it still is surely

thought-provoking. What if we are all descendants of one random couple in the past, that we all share the same grand genealogy; like how we, a group of non-prolific, not-a-dollar-to-their-name students, share Gauss as a common ancestral professor?

## 2. Define(Common Ancestor, Genealogy);

What is a common ancestor? Casually speaking, it is self-explanatory. As an example, for a randomly selected individual in a population, the individual's sibling and themselves, share a Common Ancestor (abbrev., CA) i.e., their parents. One could go further back and see that those individuals' grandparents are also a CA. Then, we can look more into this and identify that their parents are biologically the Most Recent Common Ancestor (abbrev., MRCA) to those two individuals and the generational difference is 1 in this case. We call this difference *Time to Most Recent Common Ancestor* (abbrev., TMRCA). Depending on the type of model and study we are looking at, the TMRCA can accept either a time unit (i.e. years, millennia, etc.) or number of generations (discrete natural numbers). Now, at this point, you can see that the example is a very simplistic illustration of how the MRCA would look like, intra-familial speaking. In a larger scale, the overall scientific endeavor of MRCA is to expand the scope to the human race as a whole—to find the MRCA and estimate the TMRCA of all the people living on this planet. However, the computational and resource cost of such a task is apparently too exorbitant for us to undertake, for whichever model we might use; therefore, it is more realistic, as a proof of concept, to model how a model would work by using a smaller, achievable sample population. Upcoming in this paper will also explain how MRCA can be found using various models and how TMRCA can be estimated through DNA tests.

## 3. Finding the MRCA through Biological and Computational Means

The human genetic information, or DNA, is obviously at the epicenter of scientific researches. That is, to understand many characteristics of what makes us human, we would want to further our knowledge of what *builds* human. By analyzing an individual's genome, we would know more about ourselves. It is now general knowledge that the mother's mitochondrial DNA is the main contributor to the descendant's genetic makeup, so there has been DNA tests that target this presumed distinction to estimate the TMRCA. However, because science does what it does best: breaking through the concrete textbook knowledge, we now also know that the genetic drift also occurs with the father's mitochondrial DNA. That probably means the DNA tests that rely on the maternal are flawed to some point. It is quite disorienting because one afternoon I was studying about this specific topic in Biology I, and the next day I woke up to the news that everything I had been taught was nullified by the fact that DNA now drifts paternally.

There have been many promising attempts to find the MRCA by relying on our contemporary mathematical prowess and have found that our MRCA is actually rather recent and within known human history. Even though these attempts use different models and perspectives and might even yield different results, they are all indeed very interesting to study. In biology, we fairly recently know that our genetic information (DNA) is carried through the mitochondrion (*pl., mitochondria*), an organelle in the living cell, more casually known as the powerhouse of the cell (NIH, 2019). More specifically, some specific DNAs (mitochondrial DNA, abbrev., mtDNA) are carried through the mother's mitochondria, hence the term mitochondrial Eve—referring to the matrilineal MRCA of all living humans (in simpler terms, the greatest grandmother of all living humans) (Cann et al., 2010). Because of this fact, there is such a model in case called the Wright—Fisher model. The namesakes of this model look at the maternal line in accordance to fact that DNA is transferred generationally through the mother's mitochondria and map such genetic drift to find the mitochondrial Eve lived in the past. However, more recent papers have shown that our MRCA is in fact even more recent than previously suggested and there has been proof that the father's mitochondria also contribute to the inheritance of genetic information, rather than the mother's alone (Luo et al., 2018). Therefore, there needs to be a better approach to mapping the ancestral tree in order to find the MRCA.

## 4. The Unapologetically Probabilistic Approach

Dr. Joseph T. Chang (1999) detailed a such model, complete with mathematical proofs, to estimate the MRCA of a set of randomly mating population. This model can be easily modelled using computer simulation, compared to the genetic approach of previous methods. The model that Dr. Chang suggested also consider the fact that mtMRCA drifts both maternally and paternally (both through the mother and the father) – this gives birth to the biparental MRCAs in Chang's model. For a population size $n$ and considering that all the individuals in that population engage in random mating, Dr. Chang proposes that the probability for choosing two parents from the previous generation is 1/n and have each individual in the current generation choose twice, in contrast of only choosing once in the Wright-Fisher model. Now, there are some limitations on this model. Chang's model proposed so far is an over-simplistic view of how real life would work and he pointed that fact out in his paper, assuming many effectors, i.e. geographical, racial, socio-economical, etc., to be convenient for the simulation later on. However, it is to be noted that, a novelty in Chang's model is that each individual is considered a monogamous couple; therefore, when choosing, each person in that couple would choose one couple of

a parent from the previous generation, resulting in the "choosing twice" operation per individual couple.

To prove his model, Dr. Chang (1999) employed many probabilistic functions, e.g. Poisson and binomial distributions being used in many of his lemmas and propositions. To understand the underlying mathematical complexity of his theorems, a higher level of understanding of math is to be needed. Therefore, we shall not dive deeper into them. But what we do know is that individuals in the previous generation, denoted as $G_n \mid n \epsilon N^*$, are chosen by individuals in the current generation with the probability of 1/n. According to Chang, the TMRCA follows a logarithmic function in terms of the size of the population. So, as we sample a bigger and bigger population size, our TMRCA would approach the value that of $\lg n$.

That being said, we now return to the question of whether or not we are able to reach the maxima for TMRCA using the limited computational resources to calculate and/or simulate the model. Chang (1999) also did some simulations with regards to the population size $n$ and he refers to $500 \leq n \leq 4000$ as small. Using this information, I would like to suggest that a more tantamount value for $n$ when I, myself, undertake the task of simulating the model would be around $5000 \leq n \leq 10000$ any bigger would be lengthily impractical. Though, it is to be noted that at this collegial-level course, the simulation would be rather simple in terms of programming skills and does not take into account the time and space complexity of the program. What we have so far is a purely theoretical and mathematical simulation of a population size $n$ to find the MRCA and the TMRCA. This simulation is not a realistic approach to estimate the TMRCA because of the reasons mentioned above and should only serve as a point of discussion on the premise of human history. Below is an illustrative picture of how the model simulation would work in practice for a population size of n = 6.

```
5. Help(MRCA_figure);
```

You can see that we are judging a couple as an individual in a given generation. For the present generation ($gen(0)$), we inscribe that each individual has a descendant of themselves and only themselves; thus, is given a number from 1 to $n$. Each individual in the present generation shall choose twice for their parents in $gen(-1)$ with the probability of $1/n$. And, as Dr. Chang lays out (Chang, 1999), we will also need to account for the individual only choose 1 parent in both two choices. That is illustrated by such individual only has 1 arrow coming out of that individual. With regards to gen(-1), the descendant(s) of each of the individual in gen(-1) is (are) denoted by the number of which the individual in gen(0) has in the initialization. At this point, we shall disregard gen(0) as our current generation and move on to gen(-1). We repeat the parent choosing process from gen(-2) for gen(-1) as it was done for gen(0). We can see that in gen(-2) there are two new concepts emerge from *Fig. 3*, i.e. the MRCA and extinction. The MRCA is marked with a circle and the extincted (having no descendants) is marked with a black shading. However, it is to be noted that though the MRCA is unique, it is part of a CA set. ($MRCA \in \boldsymbol{CA}$). In my illustration, the MRCA appears in gen(-2), so the TMRCA is 2. As displayed in the figure on the next page, more and more CA's appear as we go further back in time until a point where every individual in a generation is a CA of all present-day individuals.

**Figure 1.** Couple as an individual in the model

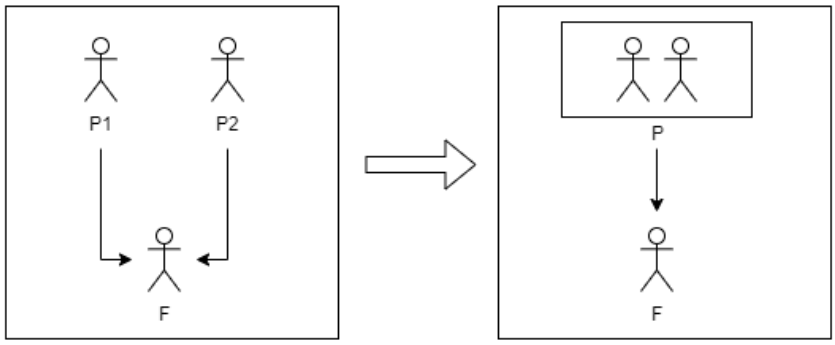


**Figure 2.** Demonstrating the Genealogy Tree based on Dr. Chang's model

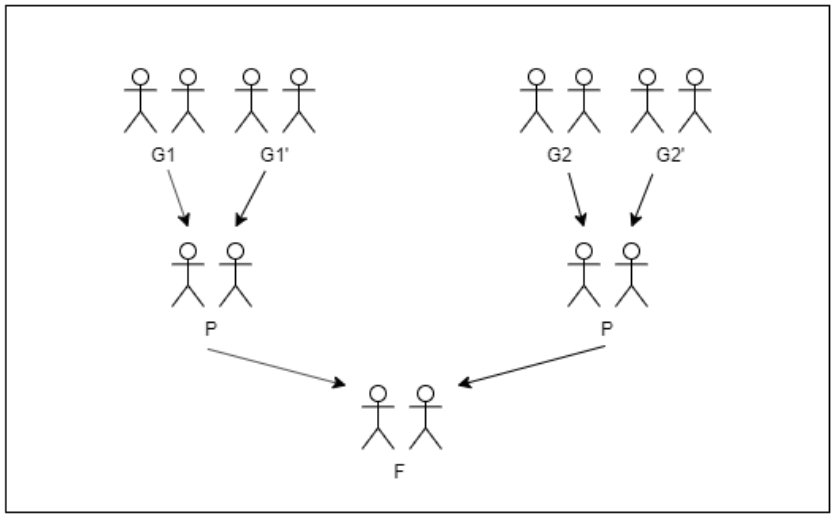


**Figure 3.** Illustration of the Simulation Model for Population Size of n=6 and the Descendant Generations Choosing Parents Twice with Probability of 1/n

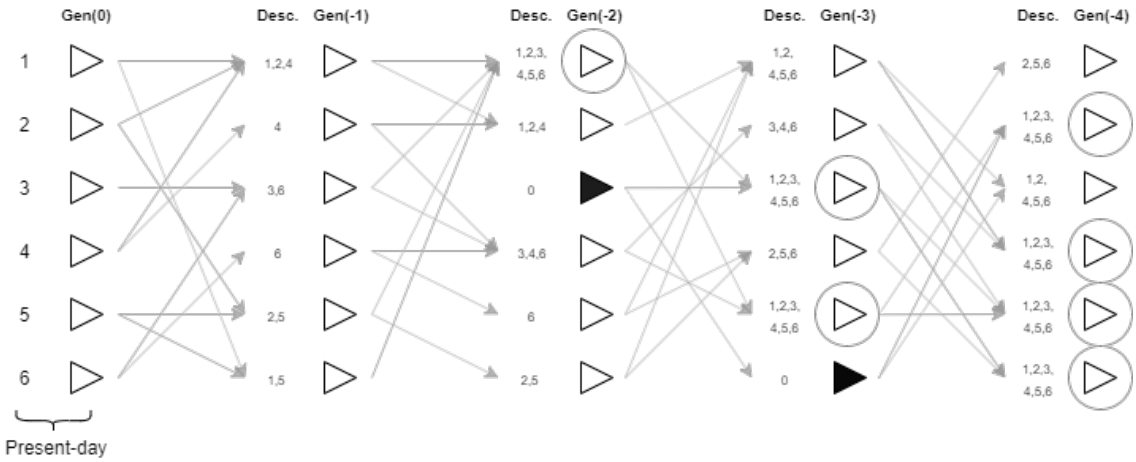


*Illustration originally created by Nguyen Hoang Anh (hoanganh.theodore@icloud.com) based on Dr. Joseph T. Chang's paper (Chang, 1999) on MRCA as part of a college project at John Abbott College.*

## 6. Estimating the TMRCA through Genetics

There are many other methods of estimating the TMRCA, some of which stem from the biological and genealogical viewpoint rather than pure mathematical like Dr. Chang's that we are discussing. There is a study by Zhou and Teo (2016) that lists different methods of estimating the TMRCA (CoalHMM, DADI, GPho-CS, MIMAR, T-FST, and T-LD), and then pits their accuracy against each other. The study finds substantial differences in accuracy even in between these genome-based methods. What does this fact tell us, the everyday Joes and Janes? Well, it could infer that even the seemingly most scientific methods may not yield the best results. So, comparing such *realistic* models to Chang's (1999) purely mathematic and probabilistic is like comparing oranges and apples – both methods will try to satiate our curiosity and fascination for the past.

## 7. Conclude();

In all, the endgame that the scientific community wants is to build a time machine (or a model that acts like a time machine) to the past and use the modern, contemporary lenses imbued with the advanced scientific knowledge to study how the human race has evolved as a whole. These models that try to find the MRCA and estimate the TMRCA (Dr. Chang's (1999) included) all serve that purpose, all have their peculiar quirks in varying scientific perspectives, and all yield various results. It is safe to say that these quirks and results are rather arbitrary in the realm of precision simulation but very crucial in our advancement into an increasingly technologically advanced society. What is gained by determining such things, such data? I really like this question because in the realm of machine learning and big data, the most important question is what meaning the numbers imply and how we can act on these numbers. So, what is ultimately gained from knowing the MRCA and TMRCA?

## References

CANN, R., STONEKING, M. & WILSON, A. Mitochondrial DNA and Human Evolution. Nature 325, 31–36 (1987) doi:10.1038/325031a0.

CHANG, J. T. (1999). Recent common ancestors of all present-day individuals. Advances in Applied Probability, 31(4), 1002-1026. doi:10.1239/aap/1029955256.

CHANG, J. T. et al. Modelling the Recent Common Ancestry of All Living Humans. NATURE, VOL 431, 30 SEPTEMBER 2004. www.nature.com/nature. https://doi.org/10.1038/nature02842

Genetics Home Reference. U.S. National Library of Medicine. U.S. National Institute of Health. Mitochondrial DNA. https://ghr.nlm.nih.gov/mitochondrial-dna. Accessed on November 5, 2019.

International Society of Genetic Genealogy. Time estimates. Most Recent Common Ancestor. ISOGG.ORG. https://isogg.org/wiki/Most_recent_common_ancestor. Accessed on November 3, 2019.

LUO, S. et al. (2018). Biparental Inheritance of Mitochondrial DNA in Humans. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.1810946115.

Mathematics Genealogy Project. Department of Mathematics, North Dakota State University. https://genealogy.math.ndsu.nodak.edu/index.php.

ZHOU, J. AND TEO, Y. Estimating Time to the Most Recent Common Ancestor (TMRCA): Comparison and Application of Eight Methods. European Journal of Human Genetics. NATURE. https://www.nature.com/articles/ejhg2015258. Accessed on November 5, 2019.