Bi/Ge105: Evolution
Homework 3
Due Date: Wednesday, January 29, 2020

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."- William Thomson (Lord Kelvin) (Popular lectures and addresses, Vol. 1, "Electrical Units of Measurement", 1883)

1. A feeling for the numbers in evolution continued

The processes of evolution take place at many different scales in both space and time. Like last week, the goal of this first problem is nothing more than to "play" with some of the characteristic scales associated with a broad range of processes in evolution. These estimates are intended to be done using simple arithmetic of the "one-few-ten" variety (i.e. few times few is ten) and to give an order-of-magnitude picture of the phenomenon of interest. Take pride in your results and state and justify (with citations) the assumptions you make carefully and give a simple, intuitive description of how you came to your results. Please don't report rough estimates with long lists of "significant" figures.

(a) This homework focuses on the role of mutations in evolution. In this part of the problem, we are going to construct our own version of an analogy with DNA replication conceived by Tania Baker of MIT. Use the Bionumbers website to get a sense of the speed of DNA polymerase in bacteria and its associated error rates during the replication process. The key assumption of our analogy will be to imagine that the diameter of DNA is scaled up to 1 m. Given that scaling, what size truck will the replication machinery correspond to? What will be the speed of this truck (in km/h and m/s)? If we think of each new nucleotide added to the newly replicating genome as a package delivered (to continue with the FedEx truck analogy), what will be the spacing between (in meters) deliveries as the truck drives and drops off packages? How long will the daily delivery run take (i.e. to complete

the DNA replication/package delivery process)? How many days between a mistaken delivery given the measured error rate in DNA replication?

2. Mutation rates by the numbers

Comparing genetic sequences has served as a useful tool for determining how various organisms are related to each other. With the advent of the "genomic era," we no longer have to infer how living organisms are related to each other based on morphological traits alone. In this problem, we will begin to get a sense of the time scales over which mutations accumulate in genetic sequences and how we can use these mutations as a molecular clocks for determining the relationships between various organisms. To get us started, we consider two different molecular "clocks" here: one for genomic DNA and one for mitochondrial DNA.

Mutations in genomic DNA.

To avoid complications of recombination that arise as a result sexual reproduction, we will only consider mutations as they accumulate in the Y chromosome. Note, since the Y chromosome maximally occurs as a single copy, it never has a chance to recombine and is simply passed down directly from father to son.

(a) Given that the human Y chromosome is around 60 million basepairs long and genomic DNA is replicated with an incredible fidelity of only one error in every 10^{10} basepairs per replication, how many mutations do you expect to see after one genome duplication? Alternatively, how many genome duplications do you expect are needed for one mutation to appear in the Y-chromosome?

With this number of mutations per genome duplication in hand, we can next tackle how many mutations are passed on from a father to his son. Recall that while many mutations may occur in a given human, only those that accumulate in the gametes (egg and sperm) will actually be passed on. To determine the number of mutations that we expect to be passed on, we will need to consider the precise mechanism by which sperm are formed, a process known as gametogenesis (see Figure 1).

As a primer for thinking about gametogenesis, let's briefly review the difference between mitosis and meiosis. Mitosis is the process by which a somatic cell duplicates its genome and then divides into two cells. Thus

in a human, mitosis yields two cells with 46 chromosomes each. Meiosis, however, is the process by which a cell duplicates its genome and then proceeds to undergo two cell divisions, ultimately resulting in four cells with 23 chromosomes. This means that each round of mitosis requires one genome duplication and each round of meiosis requires one genome duplication (despite having two cell divisions).

When considering spermatogenesis, it's important to note that this process occurs continually throughout a male's lifetime upon reaching sexual maturity (i.e. puberty). At a bare minimum, a developed sperm cell has undergone 34 rounds of mitosis (30 leading to the formation of the stem cell and 4 after the stem cell) and 1 round of meiosis. But there are also additional rounds of mitosis to take into account as the result of the stem cells continually dividing to maintain the sperm supply. With these stem cells dividing every 16 days after puberty, the number of genome duplications to make a man's sperm is dependent on the age of the man.

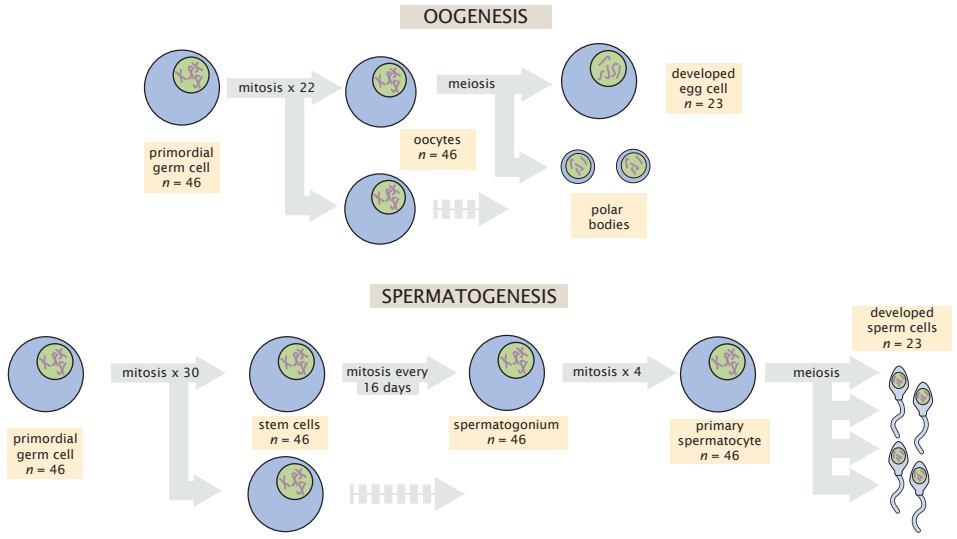

Figure 1: Schematic of oogenesis and spermatogenesis in humans. n refers to the number of chromosomes, where somatic cells have 46 and gametes have 23. For simplicity, the dashed arrows indicate the lineages of cells that we do not follow.

(b) How many genome replications have occurred to make a "typical" man's sperm? In this context, you'll need to specify what you consider to be a "typical" reproducing male. That is, what age did he hit puberty and what age does he reproduce at?

(c) Given your answer in to the previous parts of the problem how many mutations do you expect this "typical" man to pass on in his Y chromosome?

From this we have calculated the rate at which the genomic DNA clock ticks. Now we will consider a second type of DNA, whose mutations accumulate at a different rate.

Mutations in mitochondrial DNA.

Before whole genome sequencing was routinely viable, one method for comparing the sequences of different humans was to instead use their much shorter mitochondrial DNA sequence. For example, the quest to understand the human origins leading to the Out-of-Africa hypothesis (see Figure 2) was based on comparing sequences of human mitochondrial DNA. Additionally, mitochondrial DNA mutates more quickly than genomic DNA, acting as a faster molecular clock that is more useful for comparing sequences over shorter time scales. Here we explore mutation accumulation in mitochondrial DNA in contrast to the numbers to arrived at above for genomic DNA.

(d) In contrast to genomic DNA, mitochondrial DNA is replicated with much less fidelity, incurring an error rate of 3×10^{-5} mutations per base pair per generations. Given that the entire mitochondrial genome is around 17000 bp, how many mutations do you expect to accumulate in the mitochondrial DNA per human generation?

3. Comparing our molecular clocks over time

Now that we have determined the number of mutations that humans pass on with each generation, we can begin to think about how mutations accumulate over evolutionary time scales. Specifically, if we were interested in assessing how related two individuals are, we could sequence their genomes (or part of their genomes), align the two sequences, and count the number of differences. We can then use the number of differences as a metric for

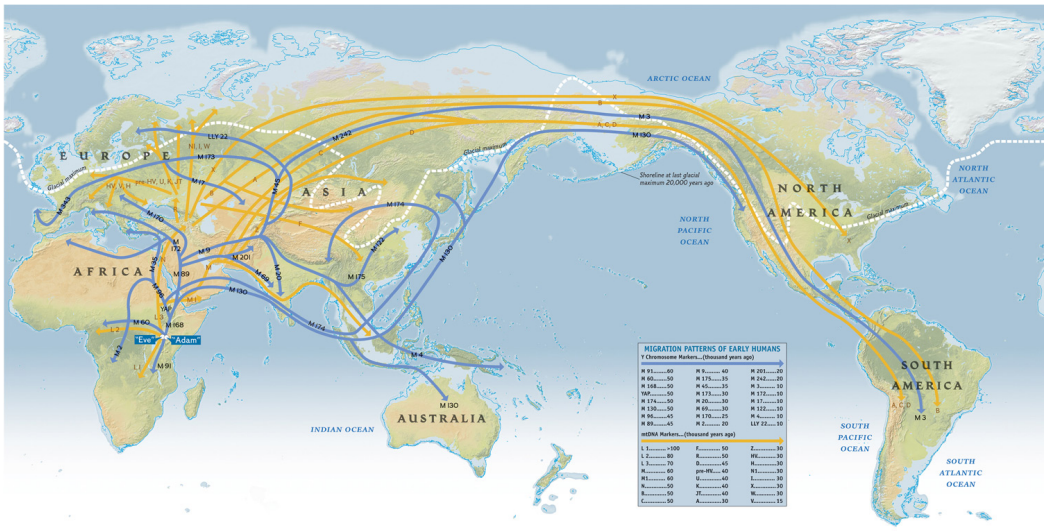

Figure 2: This map shows the patterns of human migration as inferred from modern geographical distributions of marker sequences in the Y chromosome (blue), indicating patrilineal inheritance, and in the mitochondrial DNA (orange), indicating matrilineal inheritance. Both methods suggest a similar geographical location for modern human origins in east Africa. Source: National Geographic Maps, Atlas of the Human Journey.

how related the individuals are, where more differences indicate that the organisms are more distantly related. Note that we are oversimplifying the process of evolution by only considering mutations that arise as a result of single basepair substitutions, but these estimates will give us a sense of how we expect sequences to be related over time.

(a) Some of the oldest people on Earth have lived to be around 120 years old. Imagining a somewhat futuristic scenario where everyone's genome is sequenced at birth, by how many mutations would you expect one of these 120 year-old men and his present-day male descendants to vary on their Y chromosomes? [Note: Let's assume that the male linage is continued at each generation. Additionally, you'll have to make some assumption about generation time.] While this exact situation is extremely contrived, we may not be too far from a future where everyone's genome is routinely sequenced and this type of longitudinal data could be obtained.

For contrast, let's now consider a much longer time scale.

(b) Given the age of Lucy that you determined in Homework 1 (i.e. around 3 million years), by how many mutations in the Y chromosome would you expect modern-day males to vary from a male counterpart of Lucy?

To get the estimate above, we have been working under the assumptions of the neutral theory of evolution, which states that the rate at which mutations become fixed in the population is equal to the rate at which mutations occur in the first place, as long as the mutations are neutral. This means that while we won't be able to see all the mutations that have occurred between Lucy and present-day humans, the present-day genomes and the mutations that have become fixed serve as a good proxy for the mutation rate during this intervening time. While this neutral theory of evolution may apply for large segments of the genome, it likely isn't true for the entire genome. That is, mutations in many essential protein-coding genes would not be neutral but would in fact be quite deleterious to to organism. So rather than compare two whole genomes, it might be more reasonable to compare a smaller region of the genome that we are confident is not essential to the cell. By using polymerase chain reaction (PCR) as we've done together in the lab, we can easily amplify 5000 bp of DNA from two individuals and have this smaller segment sent out for sequencing, a method that is both cheaper and faster than whole genome sequencing while also avoiding issues of violating the neutral theory of evolution.

(c) Given the number of mutations across the Y chromosome that you estimated for parts **(2a)** and **(2b)**, how many mutations would you expect to see if you were only looking at one of these 5000 bp segments?

(d) Repeat **(2a)**, **(2b)**, this time for mitochondrial DNA. That is, by how many mutations would you expect a 120 year-old woman and her present-day female descendants to differ in their mitochondrial DNA? What about between Lucy and modern humans? Do you have any concerns about the numbers you get?

With radioactive decay dating, the method we wish to use is often determined by the time scale we are interested in. For example, we may use

carbon dating for short time scales, while we would use potassium-argon dating for longer time scales as we saw in Homework 1.

(e) Drawing an analogy from radioactive decay dating: From the calculations you did above, when would you want to use mitochondrial DNA to compare the relatedness of individuals? When would would you want to use genomic DNA? Within genomic DNA, there may be variability in the rate at which different sequences accumulate mutations: what would be examples of regions that you expect to mutate more or less slowly within genomic DNA?