

Estimating the Age and Selective Advantage of the *CCR5*Δ32 Allele in Humans

Introduction

HIV, the virus that causes AIDS, is currently having a profound selective impact on the global human population. According to the Joint U.N. Program on HIV/AIDS, by the end of 2007, 33.2 million people were living with HIV/AIDS, and approximately 25 million people have already died of the global epidemic. Under these conditions, we would expect any mutation that conferred even partial protection against HIV to increase substantially in frequency.

As you have read in the paper by Stephens *et al.* (1998), in 1996 researchers discovered a 32-bp deletion in the *CCR5* gene, which codes for a chemokine receptor on many cells of the immune system. This deletion, termed the *CCR5*Δ32 allele, delays by 2-3 years the onset of AIDS symptoms in *CCR5*-+/ Δ heterozygotes. Furthermore, *CCR5*- Δ / Δ individuals appear almost completely resistant to infection by most strains of HIV. Many studies since have further elucidated the function of the *CCR5* gene and its mutant allele. In this week's lab, we will follow Stephens *et al.* by using computer simulations to estimate the selective advantage (if any) of the *CCR5*Δ32 allele.

Population Genetics

1. Mutation

All genetic variation in a population originally arises by mutation. Mutation rates are very low for most organisms, typically on the order of 10^{-6} mutations per generation for each allele. Other, more powerful evolutionary forces therefore play a much more important role than mutation in altering allele frequencies. However, it is important to recall that those other forces (such as natural selection and genetic drift) can act only on the variation that mutation has already produced. Without mutation, evolution would rapidly grind to a halt.

2. Genetic drift

Genetic drift is change in allele frequencies that is caused by random fluctuations from one generation to the next. To see how this happens, imagine that a population consists of just two individuals, with genotypes *Aa* and *Aa*. In this population, $\text{freq}(A) = 1/2$ and $\text{freq}(a) = 1/2$. Now allow these two individuals to mate. On average, the allele frequencies in their offspring will still be $1/2$ and $1/2$. However, based on Mendelian inheritance and the laws of probability, the actual frequencies may differ from those expected values. For example, the parents could produce two offspring with genotypes *AA* and *Aa*. Under this scenario, the allele frequencies in the new generation are $\text{freq}(A) = 3/4$ and $\text{freq}(a) = 1/4$. This is an example of genetic drift because the change in allele frequencies results entirely from random chance.

Genetic drift can be modeled as a "random-walk" process in which allele frequencies in each generation differ slightly and randomly from frequencies in the previous generation. Specifically, if p is the frequency of an allele in one generation, then the frequency in the next generation is sampled randomly from a normal distribution with mean p and variance $pq/2N$. Evolution by genetic drift thus occurs most rapidly in small populations. In very large populations, the variance $pq/2N$ is so small that genetic drift can effectively be ignored. In such populations, changes in allele frequency result almost entirely from natural selection.

3. *Natural selection*

Natural selection is change in allele frequencies that is caused by genotypic differences in survival and/or reproductive success. For example, spinal muscular atrophy in humans is a serious neuromuscular disease caused by an autosomal recessive allele a . Individuals who are homozygous for the a allele are only 10% as likely to survive and reproduce as AA and Aa individuals. The table at right shows the relative fitness W of each genotype.

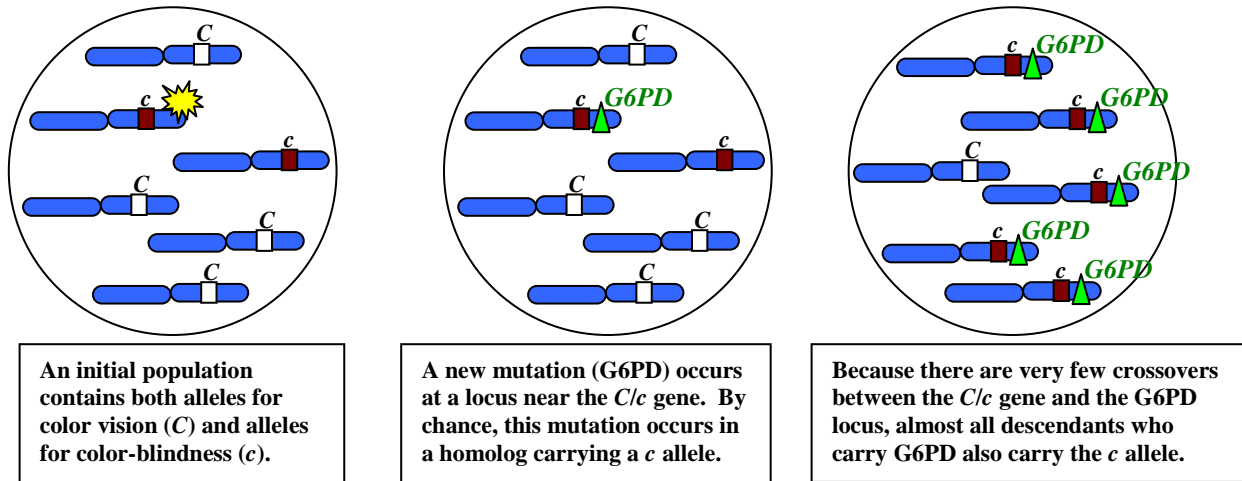
| Genotype | Relative fitness |
|-----------------|-------------------------|
| AA | 1 |
| Aa | 1 |
| aa | $1 - s$ |

Individuals with low-fitness genotypes are less likely to survive and leave offspring. Over time, natural selection tends to decrease the frequency of alleles that are associated with such low-fitness genotypes. This change in allele frequency is fastest when selection is strong. However, other factors — particularly a trait's pattern of inheritance — may also play a role.

Modeling Linkage Disequilibrium to Estimate the Age of *CCR5Δ32*

As you know, genes that are located very close together on the same chromosome are usually inherited together as a single unit or *haplotype*. For example, stubble bristles and bithorax abdomen are tightly linked in *Drosophila*. As a result, if a fly has alleles for both of these traits on the same homolog, nearly all offspring that inherit the allele for stubble bristles from that parent will also inherit the allele for a bithorax abdomen.

A new mutation such as *CCR5Δ32* is originally present in only one individual, on one of that individual's homologs. Any offspring who inherit this mutation will also inherit any other alleles that happen to be nearby on that same homolog. For example, color-blindness is very common on the island of Sardinia. This is because the G6PD mutation, which confers resistance to malaria, first occurred on a homolog that also contained an allele for color-blindness. Natural selection favored the malarial-resistance allele, which increased in frequency. However, because the two loci are so tightly linked, almost all children with the malarial-resistance allele had also inherited a color-blindness allele from the same parent. This phenomenon, in which combinations of certain alleles occur more or less frequently than would be expected by chance, is called *linkage disequilibrium*. (See figure below.)



In the Stephens *et al.* study, 85% of the sampled haplotypes that carried the *CCR5Δ32* allele also carried the *GAAT197* and *AFMB215* alleles at nearby microsatellite loci. Combining these data with estimates of mutation and recombination frequency, **they inferred that the *CCR5Δ32* allele first arose approximately 27.5 human generations ago** (95% CI: 16–31 generations).

For the remainder of this activity, we will independently infer *CCR5Δ32*'s age under a variety of population genetic models (e.g., no selection, weak selection, strong selection). If Stephens *et al.*'s estimate as reasonably accurate, then we can reject any population genetic model that yields results incompatible with that estimate. This will allow us to test specific hypotheses about the evolutionary history of the *CCR5Δ32* allele.

Estimating the Age of *CCR5Δ32* using Population Genetic Models

When a new mutation such as *CCR5Δ32* occurs, the population contains only one copy of the mutant allele and $2N - 1$ copies of the wild-type allele. Under genetic drift alone, the probability that the single mutant allele will eventually replace the wild-type allele (“reach fixation”) is $1 / (2N)$. In a population of 5,000 individuals, for example, a neutral mutation has only a 1 in 10,000 chance of reaching fixation. Most such neutral mutations are lost to genetic drift within just a few generations. Even favorable mutations are often lost to genetic drift.

Standard population genetics analyses rely on *prospective models* of a population’s allele frequencies. Such models begin with known initial allele frequencies, then use recursion equations to calculate the frequencies in each subsequent generation. (These equations may be either deterministic or stochastic.) For example, a prospective model might determine the average time needed for an allele’s frequency to increase from 0.001% to 10% under specific values of selection and genetic drift.

In some situations, however, retrospective models may be more appropriate. This type of approach underlies the field of *coalescent theory*. Coalescent models reverse the time axis to study a system’s evolutionary history. For example, a model might determine the average time an allele would have needed to reach a known current frequency of 10% from a starting frequency of 0.001%. One major goal of this lab exercise is to explore the differences between prospective and coalescent models.

Procedure: Prospective Model

1. Orient yourself to the **ProspectiveModel** simulation in *Excel*.
2. Note the initial allele frequency, genotypic fitnesses, and population size, then run the simulation. What does the result indicate?
3. Repeat the simulation several times for the same values as your previous simulation. How much and in what ways do the results of one simulation differ from those of another? (You may find it useful to extend the simulation for more than 500 generations by copying and pasting formulas on the Calculations sheet.)
4. Determine how many generations it takes for the *CCR5Δ32* allele to reach a frequency of 10%. If genetic drift eliminates the allele in a given simulation, what does that tell you about the system? How will this affect your calculations of the allele’s age?
5. Use the “Multiple Runs” macro to run the simulation at least 40 times. If possible, record the mean value, range, and 95% confidence interval of the *CCR5Δ32* allele’s estimated age.
6. How useful was this model in estimating the allele’s age? What aspects of the model, if any, most severely impacted its utility?

Procedure: Coalescent Model

1. Orient yourself to the **CoalescentModel** simulation in *Excel*.
2. Note the initial allele frequency, genotypic fitnesses, and population size, then run the simulation. What does this result indicate? How does this differ from the results of the prospective model?
3. Repeat the simulation several times. (Again, you may want to extend the simulation for more than 500 generations.) Begin collecting data on the number of generations required for the *CCR5Δ32* allele to coalesce to a single copy. How does the performance of the coalescent model differ from that of the prospective model?
4. Use the “Multiple Runs” macro to run the simulation at least 40 times. If possible, record the mean value, range, and 95% confidence interval of the *CCR5Δ32* allele’s estimated age.
5. Does the 95% confidence interval include Stephens *et al.*’s estimate of the allele’s age based on breakdown of linkage disequilibrium? What does this imply about the specific selective intensity you were modeling ($s = 0.01$)?
6. Choose other values of selective intensity s to model, and repeat steps 4–5. Try to find the precise range of selective intensities that are consistent with the age that Stephens *et al.* inferred based on linkage disequilibrium.

Overall Interpretation

- What are the strengths of each analytical method (prospective model vs. coalescent analysis)?
- What can we infer about the historical levels of selective pressure experienced by *CCR5Δ32*? To what extent do our conclusions depend on specific model assumptions such as mode of dominance, population size, and mean generation time?
- What different kinds of events could cause the levels of selection intensity you inferred, bearing in mind the geographic area and timeframe under examination? How might we go about assessing the support for those different hypotheses?