Begun Exam

Carl Boettiger

June 11, 2008

1

The basic tenets of the neutral theory of evolution are as follows: Most mutations are strongly deleterious, and thus quickly removed from the population by natural selection. Any advantageous mutations are very rare and strongly selected for, and thus quickly fixed. Thus the standing variation we see is almost all neutral. Several consequences of neutral evolution are easily derived from these assumptions, which stand at the core of tests of the neutral theory. For instance, the probability that a new neutral mutation fixes in the population is $\frac{1}{2N_e}$, where N_e is the effective population size (see question 4) while the rate at which those mutations enter the population is $2\mu N_e$, therefore mutations are fixed in the population at a rate given by the product, $\rho = \mu$, which is notably independent of the population size. The timescale for this event to occur goes as the population size, $T_s = 4N_e$, where T_s is the mean sojourn time. Because the rate of evolution (rate of divergence, rate at which mutational changes are fixed) doesn't depend on the population size in the neutral model, this is often taken as one of the first indicators for or against neutrality. The result is rather different in the case of selection, where $\rho = 4s\mu N_e$, (where s is the selective advantage of the mutation). That is, under selection ρ increases linearly with population size while the sojourn time is $T_s = \frac{2}{s}\ln(2N_e)$, which increases only very weakly with population size.

One method of detecting a recent adaptive sweep, Tajima's D, is described in Question 2, which compares measures of heterozygosity. The expected heterozygosity under a neutral model will be much higher than under a selective sweep, as described in Question 2. However, recent rapid population growth will produce the same signal, as it will in other methods,

such as measures of linkage disequilibrium, LD. The MacDonald-Kreitmann test is another method, and one that is more robust to demographic variation than many others. The MacDonald-Kreitmann test, diagrammed in Figure 3, considers a population of related individuals (i.e. samples from the same species) and an out-group comparison (i.e. its sister species). The test partitions differences along the DNA sequence of an allele in two ways: whether the difference is polymorphic in the focal population or a fixed difference between the focal population and the out-group, and also whether the substitution is silent – changing the DNA sequence without changing the protein for which it codes, owing to the redundancy of the genetic code – or non-silent, resulting in an amino acid substitution. Under the neutral theory, non-silent and silent changes are both thought to be neutral, so we should observe no difference in the ratio of silent to non-silent sites that are polymorphic compared to the ratio of silent to non-silent sites that are divergent (see figure). Meanwhile, if adaptive selection is operating on protein evolution, we expect that the silent mutations are neutral, while the non-silent are affected by selection. Consequently we expect a significant difference in these ratios, as the two categories are not comparable. A simple chi-squared test can be used demonstrate this.

2

The coalescent of a population is the point back in the genealogy at which all members of that population shared a common ancestor. The coalescent is a central concept in population genetics, and a powerful tool for understanding neutral, balancing and adaptive evolution. One useful concept associated with the coalescent is the *time in the coalescent*, T_C, which is represented by the total amount of branch length in the genealogy, see Figure 1. From this, we get a relative estimate of the number of segregating sites, $S_n = \mu T_C$. This is a measure of at how many points we expect the alleles of our 8 samples (hence $n = 8$) shown in Figure 1 to differ. From this, we can calculate our first estimate of heterozygosity of the population, $\hat{\theta}$,

$$\hat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \ldots + \frac{1}{n-1}} \tag{1}$$

Note that this depended only on the total branch length, and not the topology. $\hat{\theta}$ is an estimate of the population heterozygosity, $4N_e\mu$. Mean-

while, another measure of heterozygosity, the expected number of pairwise differences $\hat{\pi}$ between our alleles, will depend on the amount of shared history they have. Figure 2 shows a coalescent tree with approximately the same total branch length, but different relative lengths between coalescent events (each node, where two individuals share a common ancestor, is the coalescent event for that pair). Consequently, we would expect the same total number of mutations among the alleles, μT_C, but if we through these mutations at random onto the tree in Figure 2, many more of them will be shared between the first seven alleles than would have been in Figure 1. Consequently, we expect many fewer pairwise differences, $\hat{\pi}$ in tree Figure 2 will be much less than it is in tree figure 1, even though they have the same $\hat{\theta}$.

The tree in figure 1 has approximately the expected relative branch lengths of a neutral model of the coalescent (where the coalescent times are all drawn at their average time). Given eight individuals, we expect a coalescent between some pair of them relatively soon. Each time a coalescent occurs, the next coalescent is likely to take longer, simply because their are fewer nodes left to coalesce. The last two nodes can be expected to take twice as long to coalesce as the coalescent when there are four nodes, as sen higher in the tree. Under this neutral structure, $\hat{\theta}$ and $\hat{\pi}$ should agree, as they were both derived as estimates of the total heterozygosity of the population $4N_e\mu$ under a neutral model.

Meanwhile, figure 2 $\hat{\pi}$ will be much less than $\hat{\theta}$, as many of its coalescent events are much too soon. This suggests a selective sweep, where most of the population shares a common ancestor in the much more recent past than expected by the neutral coalescent, because that ancestor had been selected for. This motivates Tajima's D statistic,

$$T_D = \frac{\hat{\pi} - \hat{\theta}}{C} \tag{2}$$

where C is a normalization constant such that T_D is always between $+1$ and -1. $\hat{\pi}$ smaller than θ makes T_D negative, indicating a selective sweep, just as we intuitively expect from looking at the coalescents of Figure 2. Meanwhile, if the coalescents took much longer than expected (figure 1) then this would indicate balancing selection. Many other applications of the coalescent can be found in population genetics, but hopefully this provides the flavor of the idea.

3

The correlation between recombination rates and polymorphism is strong evidence of adaptive evolution playing a repeatably important role at many sites along the genome. Consequently they imply that a neutral equilibrium model will underestimate the population size, as the overall polymorphism will be lower than expected under the neutral model due to recent selective sweeps. High polymorphism is accompanied by high recombination, as recent selective sweeps remove polymorphism and leave sequences in linkage disequilibrium. Sections of the chromosome with higher recombination rates can more quickly break down that linkage disequilibrium and restore heterozygosity.

4

The effective population size cares much more about the history of the population than it does the current population size. Though the current population size of humans is in the billions, this number is very recent, since the human population has grown rapidly. It has been shown that for a rapidly growing population the harmonic mean population size provides a much more accurate estimate of the effective population size. The harmonic mean is always lower than the arithmetic mean (by Jensen's inequality), as it weights small numbers heavily. The small effective population size reflects that in our past, the human population was much smaller. Mutational rates have not been able to keep up with this rapid growth, and hence the polymorphism is lower than expected for a population of billions of individuals. Because of this alone, this observation is not inconsistent with population genetics theory.

There are several other explanations, including recent selective sweeps, that would also lower the heterozygosity below what is expected under the neutral model used to make this estimate, $H = 4N_e\mu$.

5

I would focus on the role of non-coding DNA. While studies have consistently emphasized that non-coding sequences are often found to be much more highly conserved than would be expected given standard estimates of

mutation rates, etc., the functional importance of these sequences has been largely undetermined. One particularly interesting candidate role is that of CIS-regulatory elements – upstream regions of the DNA responsible for turning the gene on and off. We know that gene regulation plays an essential role in many biological processes. Most obviously, almost every cell in the human body has a complete copy of the genome – meaning that liver cells and brain cells are working from the same genome. Since these cells have very different functions, they must be using that genome in a very different way, a way controlled by regulation. However, regulatory elements have much more dynamics roles as well, as gene regulation has been observed in *E. coli* to play an important role in everything from changing its metabolism to its locomotion.

First, I would identify the amount of polymorphism and divergence found across the genome, and also as grouped in coding sequences (a standard for measuring adaptive and balancing evolution in my genome), replicate pseudogenes (my best candidates for neutral evolution), and all known regulatory sequences (the focal sequences, whose statistics I aim to compare to the other two categories). I would perform a McDonald-Kreitmann test in each category to estimate the amount of adaptive evolution seen in each. I would also look for evidence of a recent adaptive sweep associated with a regulatory region by measuring the amount of linkage disequilibrium about those sequences, associated heterozygosity and Tajima's D.

In a totally different vein:

I would investigate the population genetics of viral DNA and horizontal transmission of genetic elements. I would BLAST search the complete genome of the organism and its relatives for any virus DNA known to infect my organism. I would then focus on these sections of the genome, including the genes that immediately flank the viral DNA in each genome. First, I would be able to determine if the viral DNA inserts into the same region of each genome. I would be interested to know how the amount of polymorphism and divergence in these sequences compares to that found in both coding and non-coding (preferably pseudo-genes again) regions, and also how these statistics compare in the genes immediately flanking the viral DNA to other locations in the genome. Is there any evidence of horizontal transmission of neighboring genes to the viral DNA? Does the viral DNA appear as a mutational hot-spot in the genome?