Additive alleles and drift Part I, or how to be a bad TA

Silas Tittes

August 19, 2015

History

My 3rd semester exam reading list includes *The origins of theoretical population genetics* (Provine, 2001), a fantastic little text that provides a historical account from Darwin up to the modern synthesis. One of the central ideas Provine returns to frequently (and rightly so) is the conflict between the biometricians (Karl Pearson) and the Mendelians (William Bateson) concerning the nature of inheritance of phenotypes and the resulting dynamics of evolution and natural selection: does evolution happen continuously with slight modifications over time, as Darwin had proposed, or does evolution occur in larger discontinuous jumps, as the rediscovery of Mendel's principles seemingly suggested? Depending on where one looked, support for both the biometrical and Mendelian points of view were supported. For example, when short and tall plants of the same species are crossed, any one offspring might be intermediate in stature, and looking at all the offspring, one might observe a spectrum of heights spanning the two parent heights. However, one could also observe from the same cross that the flower petals of the offspring were always one parent color or the other, but not intermediate, as with the heights. Was there truly a pluralistic duality at play, or was there some unknown unifying explanation? The answer is the latter, and as Provine explains, the scholarly contempt that the biometricians and Mendelians held for each other stalled the solution for something like 15 years. It wasn't until R.A. Fisher, with all his mathematical prowess, that we now have a satisfying answer given in his 3,000+ cited paper, *The Correlation between Relatives on the Supposition of Mendelian Inheritance* (Fisher, 1919).

Frankly, I haven't read Fisher's paper in any depth, and I don't know how much I could understand if I did. According to the Wikipedia page about the article, Punnett and Pearson refereed the paper, but had "reservations" about accepting it because of their own lack of expertise. If they didn't get it, I don't hold out too much hope for myself at this point in my education (I'll get there. We'll all get there). None the less, I wrote this blog post to explore how the many alleles, assorting independently and making an additive contribution to a single trait, can result in a near-continuously distributed gradient of phenotypes. As usual, I'm awestruck at the ease of doing simulations in R (I often wonder how Fisher, Wright, and Haldane would have felt about the degree of modern reliance on computers for our work, and how much they themselves would have partaken).

Bad TA

I recently finished TAing Genetic for the summer. We discussed continuous traits that are a consequence of many loci acting additively. We discussed the equation, $p = 2n + 1$, where p is the number of phenotypic classes and n is the number of loci. I did my best to explain this equation by making a branching diagram and recording a table of how many more unique phenotypes there are for each locus added. This went okay, but I didn't feel this really got the point across, so for funsies, I tried explain using R. Without fail there is at least one student who loves R. She will get something out of the demonstration, but the other 80-90% of the class will role their eyes and space-out. My department has really pushed to make using R one of the take-home skills for undergraduates. There's still a lot of reluctance, but coding is hip and students (myself included) are realizing it's a very marketable skill.

My code has gone through several iterations. What I presented to the class was very different from what I have below. This code is the fastest, and most effective demonstration of using R (or at least as effectively as I can use it). The speed and brevity of the code below has two costs: connection to biological process, and flexibility. By making it run quick and relying a little more on math, the code doesn't obviously resemble the process of sampling gametes, so is less educational for students interested in the biology, but are unfamiliar with some basic probability. Secondly, some simplifying assumptions have to be made, which I'll explain more below, and these assumptions make it harder to explore when and how exceptions arise, and how this process changes over time due to sampling error (drift). I will write another blog post that explores these ideas... one day.

Funsies

I have simple goal in mind – connect Mendel's principles of inheritance that apply to individual loci, and demonstrate how many independently assorting loci result in a near-continuous gradient of phenotypic classes. I've separated this into two parts: determining the number and frequency of unique trait values, and determining what the trait values actually are. For this post, let's assume each locus that contributes to the trait of interest has two alleles that add a different amount to the overall phenotype, and that the same two alleles "types" occur at every locus. Finally, as mentioned, each locus assort independently from all the others – as if each locus were on a separate chromosome.

There are many possible combinations of alleles at each locus. If two heterozygous individuals are crossed, from Mendel's rules at a single locus with two alleles (let's call them R and r to be weird), we know the genotypes and genotypic frequencies are:

$1/4RR$

$2/4Rr$

$1/4rr$

Now let's assume that the alleles, R and r, correspond to a quantitative value that contributes to a quantitative trait like height. We'll say, R contributes 10 units and r contributes 1. So now the frequencies and values of each phenotypic class are

$1/4RR = R + R = 10 + 10 = 20$

$2/4Rr = R + r = 10 + 1 = 11$

$1/4rr = r + r = 1 + 1 = 2$

With more loci, there are of course more possibilities and more variations in the phenotypic classes because we add each allele's value to the overall phenotype, which is where I turn to now. An intuitive way to think about it, as long as there are only two alleles, no matter how many loci there are, there will only be one way to make each of the two totally homozygous genotypes. Now, think about how many ways you can make a genotype that is completely homozygous except for one allele at a locus. Then think about the same, but with two exceptions to being homozygous, then three, then four, and so one until you run out of exceptions and you've reached the other homozygous genotype. If you are familiar with combinatorics, statistics, or probability, you've probably been introduced to the binomial distribution, and with it the binomial coefficient, and hearing the description above (adding another exception and so on), should remind you of the binomial coefficient, which describes the total number of ways to get a particular combination of binary events. We can use the binomial coefficient to tell us the expected frequency of each genotype given any number of loci as follows:

$p_i = \frac{\binom{n}{x_i}}{\sum_{i=0}^{n}\binom{n}{x_i}}$

where,

$i = \{0, 1, 2, ..., n\}$,

which can be thought of as the number of R alleles for the overall genotypic class p, and that all contribute the same amount.

Further,

$\binom{n}{x_i} = \frac{n!}{x_i!(n-x_i)!}$

and n is the total number of alleles, or 2 times the number of loci.

The binomial coefficient provides the frequency of each possible genotype, and we can also calculate each of the unique quantitative phenotypic values that result from the genotypes. Since we already know the frequency that each genotype will occur, we don't need to track every way to get a particular genotype, we just need to get every unique genotype, which will give us the unique phenotypes. To get each unique phenotypic class, we can us the following:

$P_i = i(R) + (n - i)(r)$,

where again, $i = \{0, 1, 2, ..., n\}$

This says we simply need to know how many of each allele type is present in a given genotype to calculate the phenotype, and that each class will be. if $i = 0$, then all the alleles of the genotype are type r, and we can multiple the number of r alleles present to get the phenotype. We know very well that the multiple alleles that contribute to a quantitative trait do not contribute the same amount. Often times we find a few loci of large effect and many more loci of small effect. The main reason for making these assumption is simplifying the math,

while still maintaining the necessary ingredients to get a meaningful result (arguably). As mentioned, I will explore the fun stuff in the next blog post.

Most the hard work is done, now it's just a matter of translating the simple math into R:

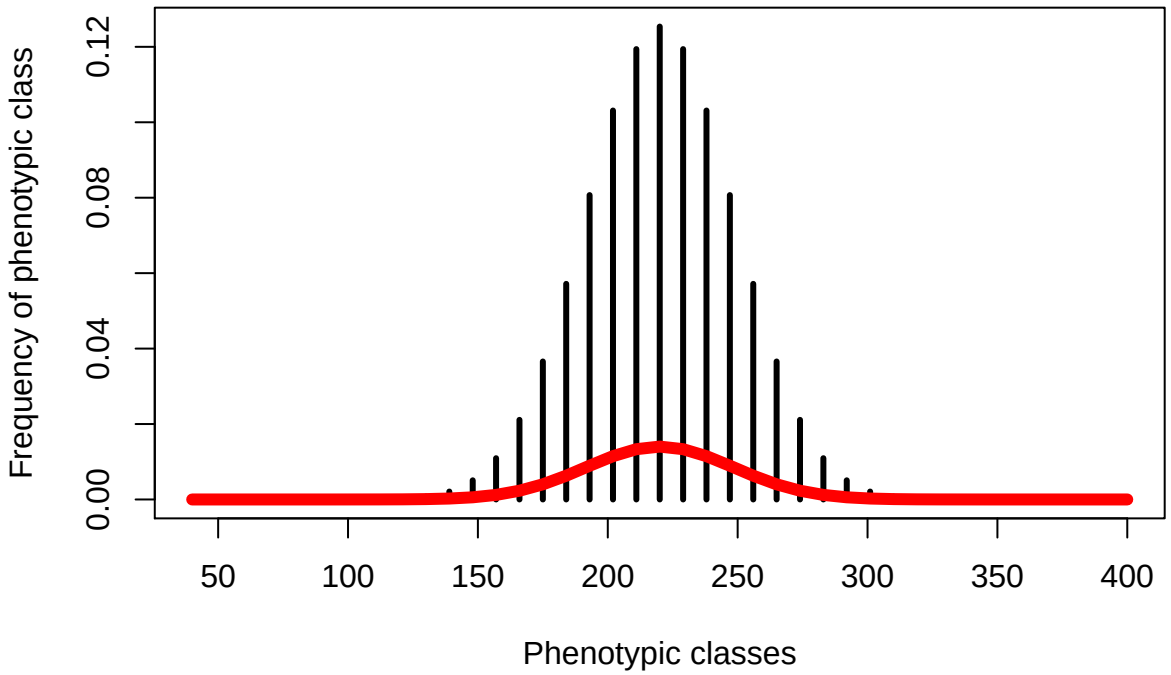
```
pheno.dist <- function(n.loci,
    allele.types) {
    n.alleles <- 2 * n.loci  #2 alleles for each locus
    # calculate the frequencies of
    # each genotype binomial
    # coefficent, starting with
    # zero copies of R, going to
    # nothing but copies of R
    class.counts <- choose(n.alleles,
        0:n.alleles)
    # turn counts into proportions
    class.freqs <- class.counts/sum(class.counts)

    # calculate the phenotypic
    # value for each genotype
    class.types <- NULL  #storage
    for (i in 0:n.alleles) {
        # from no R alleles to all R
        # alleles number of R alleles
        # times value of R + number of
        # r allels times their value
        type <- i * allele.types[1] +
            (n.alleles - i) * allele.types[2]
        class.types <- c(class.types,
            type)  #add to storage
    }
    return(list(class.types = class.types,
        class.freqs = class.freqs))
}

n.loci <- 20  #choose the number of loci
allele.types <- c(10, 1)  #choose the allelic contributions
out <- pheno.dist(n.loci, allele.types)

# visualize, gotta love that,
# type='h' option
plot(out$class.types, out$class.freqs,
    type = "h", lwd = 3, ylab = "Frequency of phenotypic class",
    xlab = "Phenotypic classes")
```

```
# See what a normal
# distribution looks like on
# top of the
avg <- sum(out$class.types * out$class.freqs)
stdev <- sqrt(sum(((out$class.types -
    avg)^2 * out$class.freqs)))
lines(out$class.types, dnorm(out$class.types,
    mean = avg, sd = stdev), col = "red",
    lwd = 6)
```



Here I used 20 loci, but really any number is possible, and it's fun to see how the frequency distribution changes as a consequence. The red line shows the would-be normal distribution of the data, which doesn't match the data all that well. I think the two reasons the normal distribution doesn't match the data are that the phenotypic classes are discrete, so that a lot of density is included in each class that would be dispersed in the infinite number of neighboring classes. Secondly, the phenotypic class distribution is truncated relative to the normal, so more that there is "too much" density near the mean. These are just intuitive guesses I'm making in place of a firmer understanding of the math. It could be that the data simply follows a different distribution, but it seems like an discrete approximate to me.

The code above runs fast, but is limited. Next time I will write about a different simulation that explores multiple loci that contribute different amounts to the phenotype, as well as the affects of drift over multiple generations. Funsies.

References

Fisher, R.A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh *52*, 399–433.

Provine, W.B. (2001). The origins of theoretical population genetics: With a new afterword (University of Chicago Press).