DNA Sequence Variation - Introduction

Contents

DNA Sequence Variation

The most basic level of genetic variation is, of course, that of DNA sequence.

Some data

As an example, we are going to use data from the acp29 locus in *Drosophila melanogaster*, Popset 11245776, downloaded as a FASTA file. We will use the `rentrez` package to access the data directly:

```
library(rentrez)
library(ape)
library(pegas)
library(strataG)
setwd("~/github/R_Popgen_With_GEOME/")
```

```
dat.srch <-entrez_search("popset",term="11245776")
dat <-entrez_fetch("popset",id=dat.srch$ids,rettype="fasta")
write(dat,file="acp29.fasta")
acp29 <-read.FASTA("acp29.fasta")
acp29
```

```
## 17 DNA sequences in binary format stored in a list.
##
## All sequences of same length: 702
##
## Labels:
## AY010527.1 Drosophila melanogaster isolate Zim2 ACP29 (Acp29...
## AY010528.1 Drosophila melanogaster isolate Zim26 ACP29 (Acp2...
## AY010529.1 Drosophila melanogaster isolate Zim29 ACP29 (Acp2...
## AY010530.1 Drosophila melanogaster isolate Zim30 ACP29 (Acp2...
## AY010531.1 Drosophila melanogaster isolate Zim32 ACP29 (Acp2...
## AY010532.1 Drosophila melanogaster isolate Zim37 ACP29 (Acp2...
## ...
##
## Base composition:
##     a     c     g     t
## 0.331 0.207 0.218 0.243
## (Total: 11.93 kb)
```

Note that at this point, we have saved the downloaded data as a local file, in FASTA format, which is the most basic DNA text format. Go ahead and find it on your computer and open it in a text editor to see what it looks like

We see that there are 17 sequences, each 702 base pairs in length. We're going to do two processing steps, one trivial and one critical:

1. As a cleanup for now, we will rename them as simply 1 through 17.

```
names(acp29) <-c(1:17)
```

2. We want to be sure they are properly aligned. To do so, we will use the add-on program `muscle` (which must be indepenently installed so that it can be called from R).

```
names(acp29) <-c(1:17)
acp.mat <-muscle(acp29) # align the data with muscle

#or alternatively, read it in from here
#acp.mat<-read.dna("https://ericcrandall.github.io/BIO444/lessons/Data/acp29_aligned.fasta",format="fas
```

Note that our data (acp.mat) is a DNAbin object, which is simply a matrix of sequences - one row per sequence and one column per position. This is a data type used extensively in ape and related packages. We can index it as such.

```
acp.mat[3]
```

```
## 1 DNA sequence in binary format stored in a vector.
##
## Sequence length: 1
##
## Base composition:
## a c g t
## 1 0 0 0
```

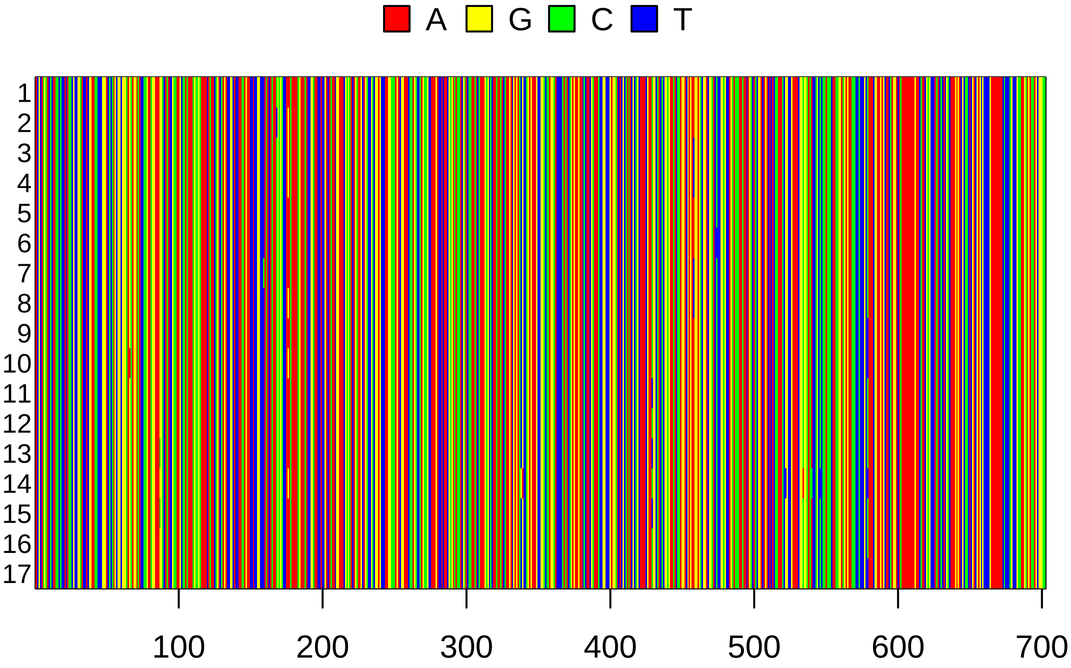
```
acp.mat[3,246]
```

```
## 1 DNA sequence in binary format stored in a matrix.
##
## Sequence length: 1
##
## Label:
## 3
##
## Base composition:
## a c g t
## 0 0 1 0
```

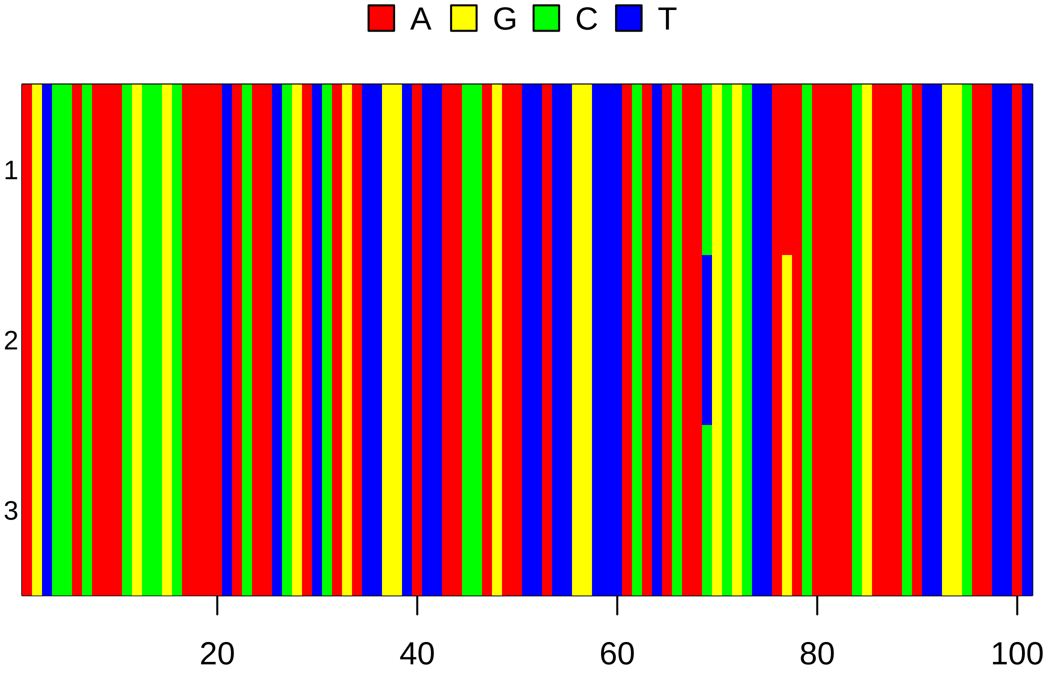
Now, we need to visualize the sequences, for a couple of reasons. First, we need to check our alignment, so that, in doing our analyses, we are as certain as we can be that we are comparing evolutionarily homologous positions. Second, we need to see, at least qualitatively, how variable the sequences are.

To do this, we will use the image.DNAbin function, part of the package "ape"

```r
image.DNAbin(acp.mat,cex=.8)
```



```r
#zoom in on the first 3 individuals, and bases 200-300
image.DNAbin(acp.mat[1:3,100:200],cex=.8)
```



And with this, we can, at least qualitatively, address our question. The fact that most bases are identical in all 17 sequences tells us that the sequence alignment looks good; the fact that there are some exceptions to this says that there is in fact *genetic variation*.

But how much variation? There are two measures that we will use extensively for this:

Number of segregating sites Since we are interested in variation, our attention should be focused on those sites that are in fact *polymorphic*, meaning that they have more than one type of nucleotide (they are therefore Single Nucleotide Polymorphisms (SNPs), or segregating sites). The ape package provides a function to give us those sites:

```
acp.sites <-seg.sites(acp29)
acp.sites
```

```
##  [1]  66  87 159 168 176 338 339 429 458 474 522 534 540 546 579
```

And we can count them rather easily

```
acp.nss <-length(acp.sites)
acp.nss
```

```
## [1] 15
```

And we see there are 15 out of the total of 702 positions that show some degree of variation. These are what we need to focus on - from a population perspective the remaining 687 sites tell us nothing.

Average pairwise differences OK, so we have 15 segregating sites. But there is another question we need to ask - on average, how many differences are there between two sequences?

Let's choose a couple of sequences at random and see how many differences there are between them.

```
seqs <-sample(1:17,2)
seqs
```

```
## [1]  6 16
```

Now let's see how many segregating sites there are.

```
ss <- seg.sites(acp.mat[seqs,])
ss
```

```
## [1] 474
```

```
ss_tot <- length(seg.sites(acp.mat[seqs,]))
ss_tot
```

```
## [1] 1
```

And we see there are 474. If we reran the above block of code, we could select two other sequences, and we'd get another number. In fact, there are n(n-1)/2 such comparisons; the average of them is our measure of diversity.

$$\pi = \sum_{ij} x_i x_j \pi_{ij}$$

Where x_i and x_j are the respective frequencies of the ith and jth sequences in the sample and π_{ij} is the number of nucleotide differences between these two sequences.

Again, there is an R function (in the package pegas) that gets at this

```
acp.div <-nuc.div(acp.mat)
acp.div
```

```
## [1] 0.004210659
```

But this is an important point to note. What we see here is the average diversity *per nucleotide*. In other words, if we were to pick a random position from two random sequences, this would be the probability that they would be different bases. We will use this occasionally, but for now, something that will be much more useful will be π, or the average number of differences *per sequence*. We can get this simply by multiplying the diversity number by the length of the sequence (702 in this case).

```
acp.pi <-ncol(acp.mat)*acp.div
acp.pi
```

```
## [1] 2.955882
```

This number tells us the average number of differences between any two sequences.

Haplotype Diversity

A final question we can ask is: how unique is any given "haplotype" in this dataset? A haplotype (think "haploid genotype") here is simply the string of bases that are found on one chromosome, but if we had data from multiple loci, it would still be all of the bases found on one chromosome.

```
acp.hap <-haplotype(acp.mat)
acp.hap
```

```
##
## Haplotypes extracted from: acp.mat
##
##     Number of haplotypes: 10
##          Sequence length: 702
##
## Haplotype labels and frequencies:
##
##    I   II  III   IV    V   VI  VII VIII   IX    X
##    4    1    3    1    1    2    1    1    2    1
```

So among the 17 sequences in the data set, there are 10 different haplotypes. Now we can calculate haplotype diversity as:

$$h = \frac{N}{N-1}(1 - \sum_{i=1}^{n} X_i)$$

Where X_i is the relative haplotype frequency of each haplotype in the sample, and N is sample size. Haplotype diversity gives us the probability that two randomly chosen sequences in the sample will be different. There is a function in the strataG package to calculate h, but we will write one ourselves:

```
hap.diversity<-function (x){
  #adapted from diversity() in StrataG
  if (!(is.vector(x) | is.factor(x)))
    stop("'x' must be a character or numeric vector, or a factor")
  x <- na.omit(x)
  n <- length(x)
  x.freq <- prop.table(table(x))
  out<- (n/(n-1))*(1- sum(x.freq^2))
  return(out)
}

acp.hap.labs<-labelHaplotypes(acp.mat)$haps
acp.h<-hap.diversity(acp.hap.labs)
acp.h
```

```
## [1] 0.9191176
```

Let’s check ourselves before we wreck ourselves by using the strataG package.

```
library(strataG)
acp29g<-sequence2gtypes(acp.mat)
acp29g<-labelHaplotypes(acp29g)
acp29g
```

```
##
## <<< gtypes created on 2022-06-20 23:02:56 >>>
##
## Contents: 17 samples, 1 locus, 1 stratum
## Other info: haps.unassigned
##
## Strata summary:
##   stratum num.ind num.missing num.haplotypes
## 1 Default      17           0             10
##
## Sequence summary:
##    locus num.seqs mean.length mean.num.ns mean.num.indels
## 1 gene.1       10         702           0               0
```

Summary

In this brief introduction, we have seen how we can quantify DNA sequence variation based on estimates of three parameters - the number of segregating sites, the average pairwise difference between sites, and haplotype diversity. However, as we shall see, there is much more that can be learned based on sequence variation, especially when it comes to making inferences about past evolutionary processes.