Sequence data in poppr: A how to.

Javier F. Tabima

February 11, 2014

So far, I guess you have used `poppr` and are aware of the multiple functionalities of this great package. Now, using these kind of data might be simple to use for a geneticist that is used to SSR, AFLP or SNP data, but what about the sequence data? We've created a couple of wrappers to ease the calculation of several statistics in population genetics such as **Tajima's D**, **Nucleotide diversity** (π) and the **dN/dS ratio**.

Now, you may ask yourself: "Why would we want new wrappers for these if they are available in excelent packages such as `pegas` and `seqinr`?". Well, because these new wrappers permit the calculation of multiple files sequentially, giving the information of several datasets that might be related (e.g. A calculation of the average π of all the genes in a genome of a species, or to define which genes of interest are under positive, negative or neutral selection).

This manual will work as a primer for the use of these functions. The data included does not represent any real dataset but will guide you through the process of using and getting familiarized with the functions. Lets begin then!

1 What is sequence data?

Sequence data can be defined as any loci, regoin of interest or DNA chunk that the researcher is interested in, as long as it has the whole information of such region in nucleotides or aminoacids. A lot of different sequence data files have been used, the most common ones being the `FASTA` and `Phylip` formats. There are many other formats that are used in different programs (such as `NEXUS, Clustal, GenBank`) but we are not interested in those. Let's see some examples of the two formats:

1.1 `FASTA` format

`FASTA` files are, by definition (Wikipedia page :P):

> "In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences."

So, the `FASTA` file has two important features: the **header** and the **sequence**. the **header** has the sequence name and could have some descriptions, for example:

```
>Sequence_Name Description
 (here is the nucleotide/aminoacid information)
```

A better example is to use an actual sequence. This example comes from **NCBI** Accession Number:XM002898690, the *Phytophthora infestans* T30-4 catalase (PITG15292) mRNA.

```
>gi|301099288|ref|XM_002898690.1| Phytophthora infestans T30-4 catalase (PITG_15292) mRNA, complete cds
 ATGGCTCCACCCACTCTTACAACGAGCAATGGCGCCCCGATGCCGCGATACGGACTGACGGCCTCCGCCACTGCTGGGTCCACTGG
 ACCACAGCTGCTCCAAGACTTTGAGTTTATCGACCACCTCTCGCATTTCGACCGTGAACGCGTCCCGGAGCGCGTCGTGCACGCCA
 AAGGGGGCGGCGCATTCGGATACTTCGAGGTGACGCACCCCGAGATCACCGACTACACGTGTGCTAAAATGTTTTCGAATGCCGGC
 AAGCGGACGCCAGTAGCAGCTCGATTCTCAATCGTGACGGCGGAATCCGGGAGCCCCGATACGATGCGAGACCCGCGGGGCTTCGC
 GCTCAA
```

In ANY fasta file, the first character that defines where the name/comments go is a ”¿” symbol. So, the name of the sequence comes after the symbol and has no spaces on it. In this case, the name of the sequence is:

```
>gi|301099288|ref|XM_002898690.1|
```

Right next to the name you have the description. Descriptions are useless for may programs but are good for researchers. In this sequence, the descriptions is giving us:

1. The species *(Phytophthora infestans)* and the sample name (T30-4)
2. The gene name (Catalase) and the category of this sequence (complete cds)

After the name/description line, the sequence data can be found. It can be nucleotide or aminoacid data. For a much more complete description of a FASTA file, go to NCBI’s FASTA file website.

1.2 Phylip format

The **great** Joseph Felsenstein developed the PHYLIP package for inferring phylogenies. In his package you needed to create a different format than the regular FASTA file. Instead of using a multi-FASTA file like this one:

```
>gi|301099288|ref|XM_002898690.1| Phytophthora infestans T30-4 catalase (PITG_15292) mRNA, complete cds
 ATGGCTCCACCCACTCTTACAACGAGCAATGGCGCCCCGATGCCGCGATACGGACTGACGGCCTCCGCCACTGCTGGGTCCACTGG
 ACCACAGCTGCTCCAAGACTTTGAGTTTATCGACCACCTCTCGCATTTCGACCGTGAACGCGTCCCGGAGCGCGTCGTGCACGCCA
 AAGGGGGCGGCGCATTCGGATACTTCGAGGTGACGCACCCCGAGATCACCGACTACACGTGTGCTAAAATGTTTTCGAATGCCGGC
 AAGCGGACGCCAGTAGCAGCTCGATTCTCAATCGTGACGGCGGAATCCGGGAGCCCCGATACGATGCGAGACCCGCGGGGCTTCGC
 GCTCAA
 >gi|301099208|ref|XM_002898650.1| Phytophthora infestans T30-4 catalase (PITG_15248) mRNA, complete cds
 ATGGCTCCACCCACTCTTACAACGAGCAATGGCGCCCCGATGCCGCGATACGGACTGACGGCCTCCGCCACTGCTGGGTCCACTGG
 ACCACAGCTGCTCCAAGACTTTGAGTTTATCGACCACCTCTCGCATTTCGACCGTGAACGCGTCCCGGAGCGCGTCGTGCACGCCA
 AAGGGGGCGGCGCATTCGGATACTTCGAGGTGACGCACCCCGAGATCACCGACTACACGTGTGCTAAAATGTTTTCGAATGCCGGC
 AAGCGGACGCCAGTAGCAGCTCGATTCTCAATCGTGACGGCGGAATCCGGGAGCCCCGATACGATGCGAGACCCGCGGGGCTTCGC
 GCTCAA
```

Phylip uses something more of this style:

```
 2 358
 gi|301099288|ref|XM_002898690.1| ATGGCTCCACCCACTCTTACAACGAGCAATGGCGCCCCGATGCCGCGATACGCTGG...
 gi|301099208|ref|XM_002898650.1| ATGGCTCCACCCACTCTTACAACGAGCAATGGCGCCCCGATGCCGCGATACGCTGG...
```

So, the Phylip format seems to be much more simpler than the FASTA format. In this case, the **header** of the file is comprised by two numbers:

1. The number of samples in the file (2)
2. The length of the sequences (358)

After the header, the following lines contain the sequence data (name and sequence). All the sequences on the Phylip format must have the same length, and are product of multiple sequence alignments.

Warning: Most of the programs that handle Phylip files only accept Phylip 3.0 format, or less than 9 characters in the name only using alphanumeric characters. For a deeper explanation of the Phylip format, go to the Felsestein’s phylip page.

2 What can we do using poppr and sequence data?

Like I said before, we've created a couple of wrappers to ease the calculation of several statistics in population genetics such as **Nucleotide diversity** (π), **Tajima's D** and the **dN/dS ratio**. lets try and give a little primer on each of these statistics.

2.1 Nucleotide Diversity (π, Nei and Li, 1979)

A simple explanation of π is that nucleotide diversity is a measure of polymorphism in a sample of gene sequences. It can be define as a summary statistic used to represent patterns of molecular diversity within a sample of gene copies. The thing about π is that can give you an approach to the gene diversity by measuring the expected heterozygosity in a single locus.

$$\pi = \sum_{ij} x_i x_j \pi_{ij} \tag{1}$$