Assignment 4.2. / Statistical population genetics 2016

Data:

Streptococcus pneumoniae bacterium, 7 gene sequences, concatenated into one sequence stretch (length 3192bp) from two population samples, Norway and USA, and from two time points, before vaccination and after vaccination. Vaccination can be expected to represent a selection pressure towards an organism.

Questions:

Are there differences in terms of nucleotide diversity and Tajima D

- Between populations (USA vs Norway)
- Are there differences between the time points

Pay special attention to the different genes in the concatenated sequence stretch. Is some of the genes especially deviating (in terms of nucleotide diversity and Tajima D) between the populations and/or between the two time points.

The genes are:
aroE 1-405 (this means that from the beginning, i.e. from nucleotide 1 to nucleotide 405, the stretch belongs to gene aroE)
gdh 406-864 (from nucleotide 406 to 864 gene gdh)
gki 865-1347
recP 1348-1795
spi 1796-2266
xpt 2267-2751
ddl 2752-3192

By using the DnaSP-option, specify these pieces of sequences as specific entities to be analysed.

It is useful to look at the differences (between populations, between timepoints) along the total sequence stretch by using the sliding window option. Note that the default window is 100/25, but it is reasonable to use a smaller window.

The files USA_before_vacc.txt, USA_after_vacc.txt, Skand_before_vacc.txt, Skand_after_vacc.txt are basicly similar population samples as are those in assignment 4.1. for human data.

In assignment 4.1. the alleles (sequence stretches) have names 1, 2 etc, which appear as codes after the mark >, and one population sample includes as many times one given allele, as appears in the frequency table.

In assignment 4.2. the allele names, or codes, are 180, 376 etc. And they are also sequence stretches (much longer than those in assignment 4.1.) One population sample, for example USA_before_vacc includes 2 x “allele 180”, 5x “allele 376” etc. (the data has been taken from one publication). So, you consider the four samples to compared, in a similar way as for assignment 4.1.