Assignment 4.1. / Statistical population genetics 2016

Human polymorphism data serves as an example for familiarizing the apportionment of genetic diversity to within-population and between-populations components on the basis of nucleotide polymorphism, Tajima D as a widely used indicator of selection-drift-demographics, and working with DnaSP-program which is a widely used and easy polymorphism analysis tool.

Human data

The data has been collected from here: http://www.pypop.org/
especially by using this page: http://www.pypop.org/popdata/2008/byfreq-DRB1.php

It might be useful to google HLA and get some all-round-education about this very interesting and important (for example, many disease associations) gene complex. HLA is an old model for genomic haplotype blocks and linkage disequilibrium, known for a long time. About ten years ago it was discovered that, in fact, the same characteristics prevail in other genomic regions, too – HLA is not an exception, it just shows the patterns very conspicuously due to it´s gene-dense structure. DRB1-gene is one of the very many HLA-genes.

The datafiles in course webpage include 9 population samples:

3 European populations: Czech, Portug, Swed
3 South American populations: Bari, Chile, Xavant
3 African populations: Camero, Oromo, Pygmi

Find answers to following questions:

- What is the the nucleotide diversity in each population and in each continent, and what are the differences between continents and between populations in general, and also between populations in each continent.

- Can something be inferred on the basis of Tajima D?

Constructing own datafiles from original data

The datafiles have already been constructed, this is an explanation how:

For example, if you want to make a population datafile from EUR Czech (see the file "HLA_DRB1_freqtable"):

This sample has 22 x allele DRB01, 21 x DRB03 etc.

From the file "HLA_DRB1_alleles" you pick up the alleles: construct a file which has

22 x

>DRB1*01
TCCTGCATGACAGCGCTGACAGTGACACTGATGGTGCTGAGCTCCCCACTGGCTTTGGCT
GGGGACACCCGACCACGTTTCTTGTGGCAGCTTAAGTTTGAATGTCATTTCTTCAATGGG
ACGGAGCGGGTGCGGTTGCTGGAAAGATGCATCTATAACCAAGAGGAGTCCGTGCGCTTC
GACAGCGACGTGGGGGAGTACCGGGCGGTGACGGAGCTGGGGCGGCCTGATGCCGAGTAC
TGGAACAGCCAGAAGGACCTCCTGGAGCAGAGGCGGGCCGCGGTGGACACCTACTGCAGA
CACAACTACGGGGTTGGTGAGAGCTTCACAGTGCAGCGGCGAGTTGAGCCTAAGGTGACT
GTGTATCCTTCAAAGACCCAGCCCCTGCAGCACCACAACCTCCTGGTCTGCTCTGTGAGT
GGTTTCTATCCAGGCAGCATTGAAGTCAGGTGGTTCCGGAACGGCCAGGAAGAGAAGGCT
GGGGTGGTGTCCACAGGCCTGATCCAGAATGGAGATTGGACCTTCCAGACCCTGGTGATG
CTGGAAACAGTTCCTCGGAGTGGAGAGGTTTACACCTGCCAAGTGGAGCACCCAAGTGTG
ACGAGCCCTCTCACAGTGGAATGGAGAGCACGGTCTGAATCTGCACAGAGCAAGATGCTG
AGTGGAGTCGGGGGCTTCGTGCTGGGCCTGCTCTTCCTTGGGGCCGGGCTGTTCATCTAC
TTCAGGAATCAGAAAGGACACTCTGGACTTCAGCCAAC

then

21 x

>DRB1*03
TCCTGCATGGCAGTTCTGACAGTGACACTGATGGTGCTGAGCTCCCCACTGGCTTTGGCT
GGGGACACCAGACCACGTTTCTTGGAGTACTCTACGTCTGAGTGTCATTTCTTCAATGGG
ACGGAGCGGGTGCGGTACCTGGACAGATACTTCCATAACCAGGAGGAGAACGTGCGCTTC
GACAGCGACGTGGGGGAGTTCCGGGCGGTGACGGAGCTGGGGCGGCCTGATGCCGAGTAC
TGGAACAGCCAGAAGGACCTCCTGGAGCAGAAGCGGGGCCGGGTGGACAACTACTGCAGA
CACAACTACGGGGTTGTGGAGAGCTTCACAGTGCAGCGGCGAGTCCATCCTAAGGTGACT
GTGTATCCTTCAAAGACCCAGCCCCTGCAGCACCATAACCTCCTGGTCTGTTCTGTGAGT
GGTTTCTATCCAGGCAGCATTGAAGTCAGGTGGTTCCGGAATGGCCAGGAAGAGAAGACT
GGGGTGGTGTCCACAGGCCTGATCCACAATGGAGACTGGACCTTCCAGACCCTGGTGATG
CTGGAAACAGTTCCTCGGAGTGGAGAGGTTTACACCTGCCAAGTGGAGCACCCAAGCGTG
ACAAGCCCTCTCACAGTGGAATGGAGAGCACGGTCTGAATCTGCACAGAGCAAGATGCTG
AGTGGAGTCGGGGGCTTTGTGCTGGGCCTGCTCTTCCTTGGGGCCGGGCTGTTCATCTAC
TTCAGGAATCAGAAAGGACACTCTGGACTTCAGCCAAG

etc., as one file (a FASTA-plain text file).

You will need the population files as separate ones, but also different populations in one merged file.
This means that you should mark the individual populations so that you can identify them while working with DnaSP which asks you to make categorizations (groups) to be analysed.

This is easily done so that when you have constructed one population, say EURCzech, you just use “edit” in notepad: replace > with >EUR_Czech. Then you have pop and continent id´s for that population. And when you merge it with other 8 populations, id-marked in a similar way, you can easily collect them for defining groups.

DnaSP

DnaSP first asks you to open a datafile: your file is not a “datafile”, it can be found by changing the window to “open all files” (because it is a textfile).

Then you should use the “Data”-window to make definitions, such as genetic code etc.

From “Define sequence sets” you can define the groups you want to analyse, all populations from one continent to one group, for comparing continents, populations separately, etc.

“Analysis” gives various analysis-options, nucleotide diversity within, between, etc. Tajima D (and other tests, which we have not discussed during the course).