Exercise 3 / BB_III 2013

During the course sessions 12.11, (19.11) and 26.11 (after the first lecture of the GWAS-module, there will not be GWAS-practicals in 26.11) we work with datasets (human~~, bacteria, virus~~) with the goal of getting familiar with widely used basic concepts of polymorphism analyses.

> Bacteria and virus datasets will be in agenda as extra exercises which you can do later, and get more than the 5 cr (which comes from the II teaching period part; note that the course continues in III teaching period for 5 ->10 cr).

Submitting reports: 5.12 at the latest => you will get personal response from teacher.

Late submissions are allowed, but might not result in response before the exam which will be 12.12.

Recommendation is that you work as groups (2-3 students) for the practical part of these assignments. However, everybody must submit an own report which is not a copy of another student´s report.

During the session 12.11 we started practical analyses and detailed instructions are now here – for those who did not attend the session.

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, working with DnaSP-program which is a widely used and easy polymorphism analysis tool, and Network-program, which is widely used for illustrating haplotype relationships as networks (network phylogenies).

Human data

Datafile HLA_DRB1_freqtable gives alleles frequencies from one very polymorphic gene, DRB1-gene from the Human Leukocyte Antigen complex (HLA), alleles DRB01, etc. from 8 continents, from 5 populations in each continent, altogether 38 populations.
Datafile DRB1_alleles gives the sequences of alleles.

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.

- For your own work choose 3 continents and 3 populations from each continent.
- What is the nucleotide diversity in each population and in each continent, and what are the differences between continents and between populations.
- For answering this, you need to construct your own files to be analysed by DnaSP.
- Can something be inferred on the basis of Tajima D?
- Analyse the population samples as haplotype networks by using the Nework program. The datafile-format needed is done by DnaSP. This was shown during the session 12.11 and details for those who work at home, explained here.
- If you don´t find enough information here, feel free to ask sirkka-liisa.varvio at helsinki.fi.

Constructing own datafiles from original data

For example, if you want to make a population datafile from the first pop which is 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).

- We look at these analyses together in our last session, 26.11 after the first GWAS-lecture .
- Try to get you files done before that.
- And, if you will not attend, ask by email for more advise.

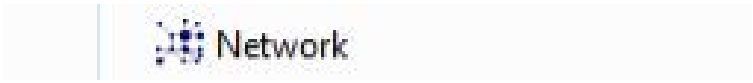
DnaSP is also a tool for creating datafile formats to be analysed by Network-software.

Here is an example by using the "HLA_DRB1_alleles"-file to be illustrated as a network. You will be using your population datafiles, instead of this, which is a kind of framework including all alleles.

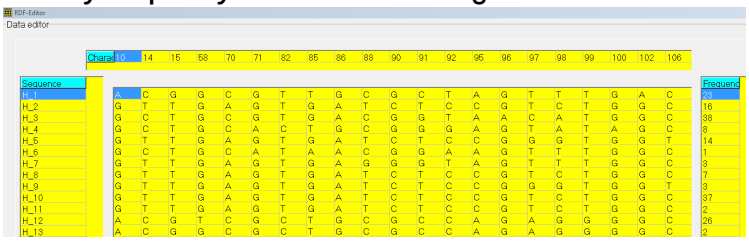
From the "Generate"-option, choose "Haplotype datafile…", and then choose "Roehl Data File (Network software)". Then save the resulting rdf-file with an appropriate name. In this example it is just "alleles.rdf", but you have, for example "Czech.rdf".

Open the Network-software (note that although it is a commercial program, the network-phylogeny part of it can be used, though not all facilities (for example, rdf-files should be done by some other program, for example by DnaSP).

From the Network-program folder you take this:



Then you pick your rdf-file and get it to the data editor window:



....in which you can, for example, give the correct names for the alleles.

Then you save the file and press exit.

Proceed to calculation menu, choose “median joining”, drawing etc. The rest is very much self-explanatory.
In fact, of course, there are several steps which we now just skip, for example parameter definitions etc. (see the attached paper and manual) – the idea now is to introduce this program – it´s usage in real scientific purposes is not just clicking something .....

What to write in your reports: to be added here and explained during 26.11 session.