Assignment 2 - BIOTECH 4BI3 Bioinformatics

When resequencing a genome a researcher is often interested in how the polymorphisms in a genome are positioned relative to the reference. They may have questions like “are the polymorphisms evenly distributed or are they concentrated in particular regions of the genome”. For this assignment, you will write a python program to parse a Variant Call Format (VCF) file and then create a polymorphism density plot from the data extracted from the file.

Briefly, a VCF file is a standard text file format to record information about polymorphisms found in a genome. The file begins with a header section (lines beginning with the ‘#’ symbol) followed by a title line with the polymorphism records appearing after that. There is one record per line and each record captures information like where in the genome the polymorphism was found, what is the polymorphism relative to the reference, and what kind of data is present to support the ‘calling’ of the polymorphism. This could be information like the number of sequencing reads supporting the call and the quality score assigned to the call. The variant call information for more than one individual can be present in a record.

To calculate the polymorphism density you do the following for each individual;

1. Establish a window of X bases wide and count the number of polymorphisms in that window
2. Record the polymorphism count and the start position of the of window
3. Shift the window down the chromosome by Y bases, count the number of polymorphisms in the window. You will be counting many of the same polymorphisms you counted in the previous window.
4. Record the polymorphism count and the current start position of the window
5. Continue moving the window down the chromosome by Y bases, counting the polymorphisms, and recording the count and position data until your window reaches the end of the chromosome
6. Do this for all of the individuals in the VCF file
7. Create a line or scatter graph of the (count, position) data for each individual. The graph should present one line for each parent (I like R but feel free to work with what you are familiar with).

The program will take the name of the VCF file, the window size, and the increment value as command line arguments.

`generate_density_data.py <vcf file> <window size> <increment value>`

The program will output to a file the data used to generate the polymorphism density plot.

This assignment will be marked on the following;

1. Correctness of function
2. Clearly written, formatted and documented code
3. Proper error handling
4. Formatting of the polymorphism density image

What to hand in:

1. A single Python program named “generate_density_data.py” which meets the requirements of the assignment.
2. An image of the polymorphism density graph using the provide VCF file.

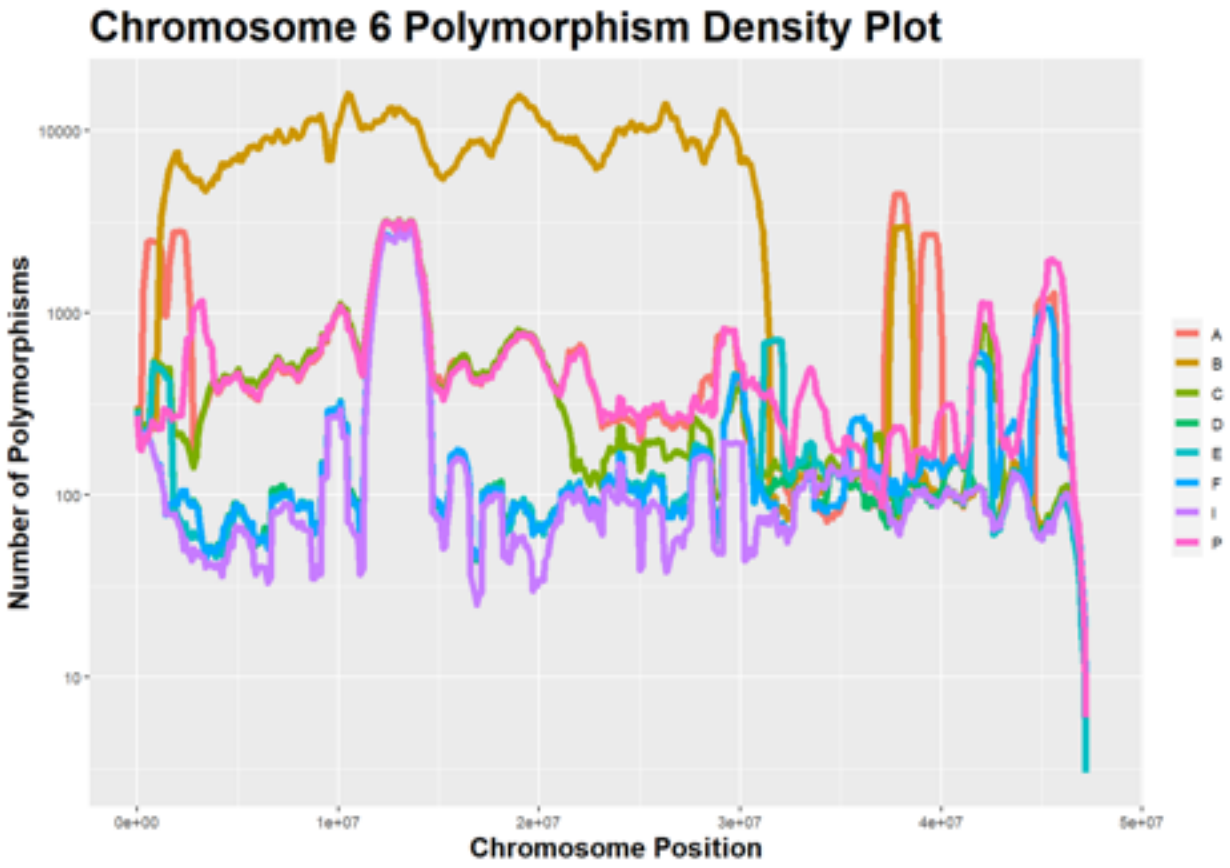

Figure 1 – Polymorphism density plot create using a window size of 1,000,000 with an increment of 100,000

HINTS:

1. The VCF file is tab delimited
2. You don’t have to use Biopython to parse the file
3. In the VCF file the polymorphism data for each individual starts with the genotype call. 0/0 means the individual does NOT have a polymorphism at that location and 1/1 means that it does have a polymorphism at that location.
4. If you save the output as a csv (comma separated value) file you can open it directly in Microsoft Excel (or R)