Genomic Analysis Services - Candidates Home Assignment

The objective of this home-task is to assess the candidate skills. It requires careful reading of the instructions, learning some basic concepts in genetic (e.g. genotyping array/DNA array), coding and creativity.
The assignment includes parsing, analyzing and extracting summary statistics of genotyping array data. You will need to understand the file formats and write dedicated code to perform specific tasks. Make sure your code does exactly what you're asked to do, e.g. create the right file formats, and that it is readable and easy to understand (meaningful comments are welcome). You may use any libraries you see fit, as long as they are publicly available. Do not worry too much about code efficiency - if your runs complete within a few minutes, that's fine.
If instructions (or anything else) are unclear, feel free to contact us by email (English) with clear questions so we can support you. The next interview will include a white board session to discuss the assignment. You will be able to sketch during the meeting. No need to prepare a presentation.

Submission

Please submit by email (to: hagar.lupo@nrgene.com; cc:kobi@nrgene.com):

- All the code you wrote (can be one or more files)
- A short explanation on how to run your code
- Output files generated by your code
- Plots/figures created by your code
- Text explaining your analyses (as requested in the instructions) and answers to questions.

Data

We'll work on genotyping array data for cotton. If you are not familiar with this technology, take some time to read about it and get the basic idea.
Links for downloading the data are provided
Additional information (as well as the data) can be found at:
https://www.cottongen.org/data/community_projects/tamu63k/data
Please note that the website may have some down time for maintenance untile August 1st.

Part 1 - array design

We'll start by parsing and analyzing the array design. Download the file TAMU_SNP63K_69997.fasta from:
https://nrgene-candidate-task.s3.amazonaws.com/TAMU_SNP63K_69997.fasta
It contains marker probe sequences in fasta format.

1. Write a script that reads the markers sequences and outputs a file with comma-separated table. Each row of the table should represent one marker and have the following information:
    a. Marker name - e.g. USDA_SNP0012
    b. Marker ID - e.g. i00001Gh
    c. Marker allele 1 - e.g. C
    d. Marker allele 2 - e.g. T
    e. Full marker sequence with allele 1 - e.g. TGCAGAACACAGA...C...AAGTAAAA
    f. Full marker sequence with allele 2 - e.g. TGCAGAACACAGA...T...AAGTAAAA
    g. Marker length - e.g. 63
    h. Position of SNP in the marker (0-based index) - e.g. 54

If markers contain more than one wildcard character, or the SNP in multi-allelic, discard the marker, but print an appropriate warning message to the screen.

2. Use the output of the above script to simplify the fasta - write a script that will read the output and print two fasta files with full marker sequences, one file for allele 1 and the other for allele 2. The fasta headers should only include the marker name.
3. Also use the table you created to get some statistics about the markers. Create appropriate figures (of your choice) to describe them in a graphical way (you may use any plotting library).

a. Marker length
b. Position of SNP in marker sequence
c. Frequencies of SNP types (C/T, C/G. C/A, A/T, A/G, G/T)
d. Any other interesting stats you can think off?

Part 2 - genotyping data

The TAMU array was used to genotype hundreds of cotton varieties. Genotyping results are provided in the file GT_AH_DiversityAnalysisDataforCottonGen.xlsx, which can be downloaded from:
https://nrgene-candidate-task.s3.amazonaws.com/GT_AH_DiversityAnalysisDataforCottonGen.xlsx

In this section, we'll parse this file and extract some statistics. You may convert the xls into a csv file for easier parsing.

1. Write a script that will read the genotyping file and outputs a file with comma-separated table. Each row of the table should represent one sample and have the following information:

a. Sample name - e.g. AH-097
b. Sample ID (entry) - e.g. 392
c. Sample description (designation) - e.g. G. amourianum
d. Fraction of markers successfully genotyped - e.g. 0.92
e. Fraction of heterozygous genotypes (out of the successfully genotyped markers) - e.g. 0.11

Ensure that genotypes match the possible alleles per marker, as defined in **Part 1**. If you find mismatches, discard the genotype and issue an appropriate warning.

2. Plot histograms of the fraction of missing data by sample and by marker. Use the results and/or any other criteria to filter low quality samples and markers. Explain the criteria you chose and create clean files (fasta and csv) with low quality markers and samples filtered out.
3. **Optional (can be also discussed as an idea w/o coding):** Create a similarity matrix, in which every cell contains the fraction of shared markers between two samples. The output should look something like this:

	Sample 1	Sample 2	Sample 3	...
Sample 1	1			
Sample 2	0.76	1		
Sample 3	0.89	0.55	1	
...				1

Which pair of samples is the most similar/different?

Note: this analysis may take a while to run on a standard laptop.

Part 3 - Create a subset array (no coding)

We'd like to offer customers an alternative array design, based on the TAMU array, that would be cheaper to run. To this end, we'll need to subset the TAMU array and choose 5,000 markers in a way that will still allow effective genotyping.
Based on the data you have, suggest one or more ways to select the markers for the new array. Explain the logic behind each method and the analyses that will need to be performed to achieve the results. Is there additional data or information that could help you make better decisions?

info@nrgene.com | www.nrgene.com