Technical take-home exercise

In the *Junior Bioinformatician* role at LabGenius your responsibilities will be to analyse and interpret raw NGS data from various biological experiments and to communicate these findings to wet-lab scientists. This task will test how you explore, analyse and interpret data, and how you communicate the results in an appropriate way for your audience.

Goal: We want to engineer a hypothetical protein with binding affinity to a clinically-relevant target using massively parallel selection assays paired with NGS. We built a sequence library composed of different combinations of the mutations of interest [**Figure 1**]. Next, we selected these variants *en masse* and enriched for variant sequences based on binding affinity. We deep sequenced the *input* [designed] library and *output* [selected] library to rank variants according to their binding affinity to our target of interest [**Figure 2**].

Materials: We've provided you with a *sequence_table.csv* file, which contains the results of the binding experiment described above and schematically depicted in **Figure 2**. This table contains variant sequences found in *input* and *output* libraries, along with the number of NGS reads corresponding to each sequence.

Task: We would like you to analyse this data in Python and produce a report of the experiment for the wet-lab biologists. Think about what information would be useful for the scientists and how it might be best presented.

To start with, we'd like you to answer the following questions:

- What do the distributions of the reads look like? Why is this?
- Do the DNA sequences match the design? Why/why not?
- What metric could you use to identify the proteins which have the highest binding affinity? Which proteins are these?

Next, we would like to see you take this analysis in whichever direction you find interesting. There are many other things you could look at in the data set, so have fun!

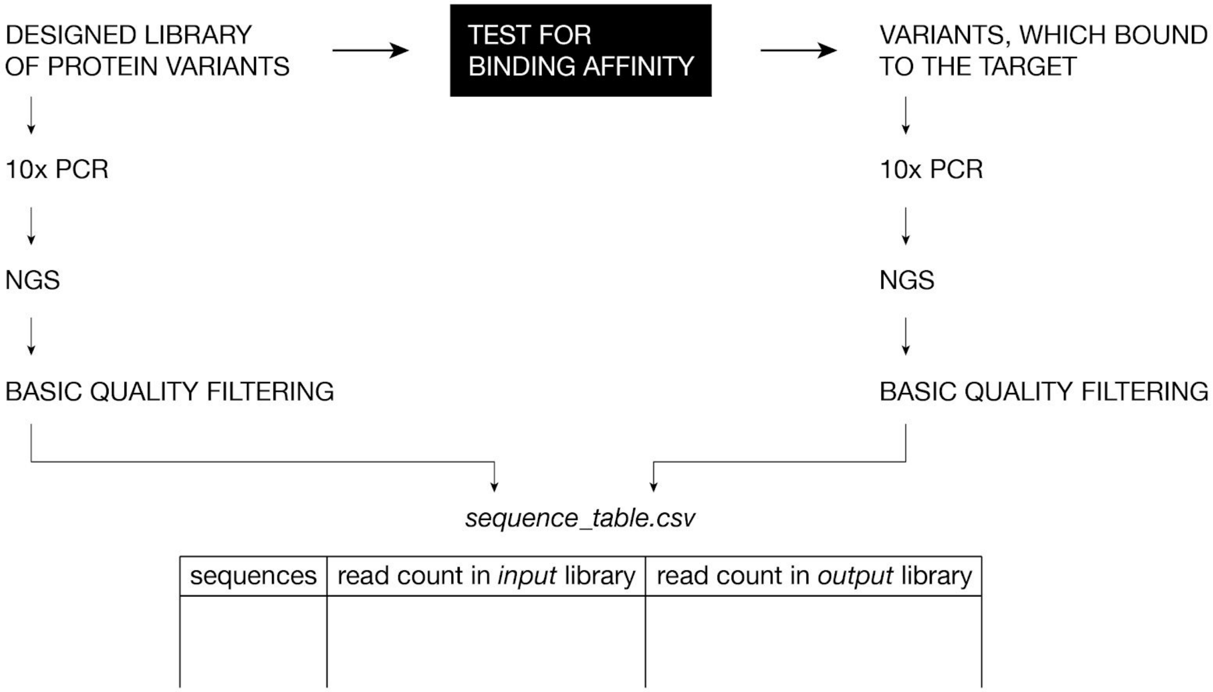
Output:

(a) A separate pdf report of your analysis and interpretation of the experimental data for wet lab biologists.
(b) Any Python code that you wrote.

Please send a zipped folder containing your files to katya@labgeni.us by the agreed time. If you have any questions about the test, feel free to send them to us within the first 30 minutes after receiving the test. Good luck!

CAGGTGCAGCTGGTGGAGTCTGGGGGAGGCTTGGTCAAGCCTGGAGGGTCCCTGAGACTCTC
CTGTGCAGCCTCTGGATTC**NNSNNSNNSNNS**TACTACATGAGCTGGATCCGCCAGGCTCCAGG
GAAGGGGCTGGAGTGGGTTTCATACATTAGT**NNSNNSNNSNNS**ACCATATACTACGCAGACTCT
GTGAAGGGCCGATTCACCATCTCCAGGGACAACGCCAAGAACTCACTGTATCTGCAAATGAACA
GCCTGAGAGCCGAGGACACGGCCGTGTATTACTGTGCGAGAGA

Figure 1. The IUPAC sequence of the variant library.



sequences	read count in *input* library	read count in *output* library

Figure 2. The experimental design that produced the *sequence_table.csv* file.