# Foresight Diagnostics

Bioinformatics Scientist, Development

The following questions aim to assess your ability to think through aspects of the types of problems this scientist may help solve. Please respond in any format you'd prefer and attach any additional code you write to solve the problem.

## Question 1

*MYD88* L265P is a common mutation in B cell lymphomas. A liquid biopsy found 5 of 1,000 molecules at the locus have this mutant allele.

(a) What is a 95% confidence interval for the mutant allele fraction (MAF)?

(b) Estimate the standard deviation for the number of mutant molecules that would be observed if more aliquots of the same sample were tested again under the same conditions.

## Question 2

Foresight Diagnostics uses phased variants to detect minimal residual disease [1]. I prompted ChatGPT: "write a python function that counts the number of unpaired reads or read pairs that have two or more mutations in a BAM file. Use type hints and add a docstring and require a minimum mapping quality". The response, saved in `count_fragments.py`, was

```python
import pysam


def count_mutated_reads(bam_file: str, min_mapping_quality: int) ->
    int:
    """
    Counts the number of unpaired reads or read pairs that have two
        or more mutations in a BAM file.

    Args:
        bam_file (str): Path to the BAM file.
        min_mapping_quality (int): Minimum mapping quality allowed
            for each read.

    Returns:
        int: The count of unpaired reads or read pairs with two or
            more mutations.
    """
    samfile = pysam.AlignmentFile(bam_file, "rb")
    mutated_reads_count = 0
    for read in samfile.fetch():
        if read.mapping_quality < min_mapping_quality:
            continue
        if not read.is_paired:
            # NM tag (edit distance) >= 2
            if read.get_tag('NM') >= 2:
                mutated_reads_count += 1
        else:
            # For paired-end reads, increment the counter only if
                this is the first in pair (to
            # avoid double-counting)
            if read.is_read1 and read.get_tag('NM') >= 2:
                mutated_reads_count += 1

    return mutated_reads_count
```

You could test it out with the provided `example.bam`.

(a) Without providing a fix, identify at least one bug in what ChatGPT wrote.

(b) Modify this function to be a script that can be called from a shell. Use any tools you'd like and assume any dependencies have already been installed. If you'd like to use another language, try changing the requested language in the prompt.

(c) *Without* implementing anything, discuss your preferred methods to ensure the accuracy and repeatability of this program as others use and modify it. Use 100 words or fewer.

## Question 3

More and more diagnostic tools are combining multiple biomarkers or measurements to understand and treat complex diseases. For example, Lymphgen is a popular classifier of Diffuse Large B-Cell Lymphomas [2].

(a) Given an assay that detects 20 mutations, you want to classify patients that respond well to a new therapy. If the classifier were the number of mutations detected, what would be the best threshold if the specificity must be $\geq 90\%$? Using the provided `data.csv` file and the number of mutations detected as a classifier, what would be the best threshold if the specificity must be $\geq 90\%$? The data has the following columns:

- `response` is the patient's response (CR = complete response and PD = progressive disease)
- `depth` is the average sequencing depth
- `mut_i` is the MAF for variant $i$. The variant is considered detected if MAF $> 0$.

(b) *Without* any additional analysis, discuss what you might do to improve the classification if you had data for 1,000 patients and were not limited to a simple count of mutations. Use 100 words or fewer.

## References


[1] Kurtz, D. M. *et al.* Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA. *Nature biotechnology* **39**, 1537–1547 (2021).

[2] Wright, G. W. *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551–568.e14 (2020).