Case study:

Determining the malaria treatment efficacy in Weveland

Please read carefully and respond to the questions below by writing short sentences/paragraphs below each question and including the relevant plots from the coding exercises. This assignment together with the associated R scripts are due from each student group by 30 March 2025. For any questions, please contact monica.golumbeanu@swisstph.ch. Many thanks for completing this assignment!

As part of the molecular surveillance expert team in Weveland, you have been charged with organizing the next Therapeutic Efficacy Study (TES) to determine the efficacy of the current malaria combination therapy used as first-line treatment in the country. Your team has already collected samples from patients that showed up at the health facilities in the past months and who tested positive for malaria.

You are using for the first time Deep Targeted Amplicon Sequencing (AmpSeq) and need to set up a workflow where you design the panel with the markers, conduct the bioinformatics analysis, and implement the infection classification and the final estimation of the treatment efficacy.

Part I: Panel design

The first key question is about choosing the right markers to interrogate in terms of their genetic variability. In order to make sure that appropriate markers are being selected, your team has already done some preliminary analyses and has selected several sensitive gene candidates. To improve the lab experiments, you would like to further refine the panel and narrow down the targeted regions to the most diverse regions of these markers. Ideally, these regions should be at most 300 nucleotides long, with high variability (high chance of mutation) across multiple positions.

First, you are setting up a workflow for designing the target regions and developing an example of procedure based on the data for one marker. For this purpose, you are using data publicly available from MalariaGEN. This data consists of mutation information for the parasites in over 20,864 patient samples. Your team colleagues have already prepared a file for you where they extracted from the MalariaGEN data the necessary information for the marker gene "*cpmp*" on chromosome 1. Precisely, for each genomic position within the cpmp gene, you have information about the reference, the alternative/mutated sequence and how many times this variation was observed in the parasites from the 20,868 samples.

Questions:

For this part of the exercise, you will need to access the *Assignments/* folder on the course page, then copy folder *Case study part 1/* to a folder in your Home directory on your computer.

1. Can you say a few words about the *cpmp* gene? It its function currently known? Do you expect this gene to be highly polymorphic (highly mutated) across parasites in patients? Why?

   - **About cpmp gene:**
     The *cpmp* gene encodes a cysteine-rich protective antigen, which is a merozoite surface and rhoptry-associated protein. This protein is located on the surface of the malaria parasite and is believed to play a role in parasite survival and adaptation.

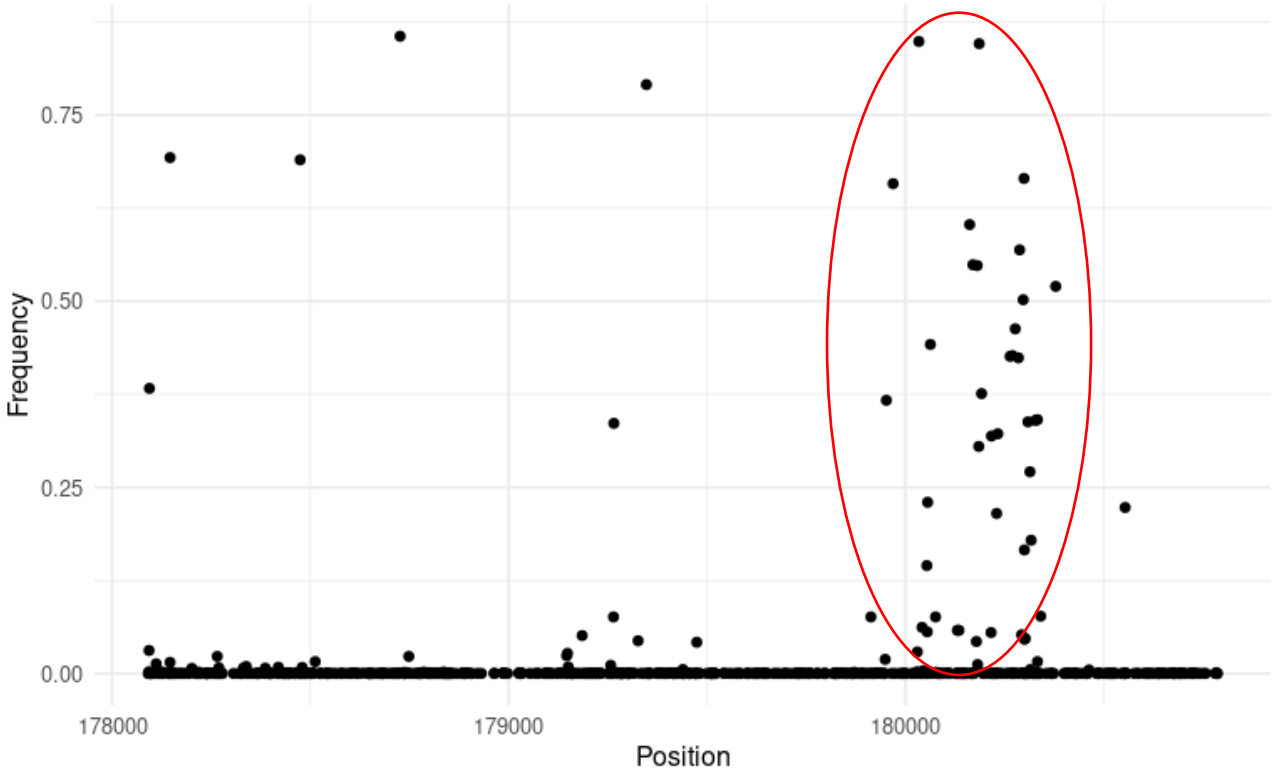
   - **It its function currently known?:**
     While the exact function of *cpmp* is not fully understood, it is thought to be involved in the parasite's response to oxidative stress or protein damage. Its surface localization suggests a potential role in immune evasion or interactions with the host immune system.

   - **Do you expect this gene to be highly polymorphic?**
     Yes, the *cpmp* gene is expected to be highly polymorphic across parasite populations.

   - **Why?**
     As a surface protein, *cpmp* is exposed to the host immune system, making it subject to strong selective pressure. This pressure drives the accumulation of adaptive mutations, allowing the parasite to evade immune detection. As a result, the gene exhibits high genetic diversity, with variations observed across different parasite strains.

2. The file *cpmp_malariaGEN_mut.txt* contains at each position on the *cpmp* gene the information about the observed variation in the malariaGEN samples (as described above). To assess the diversity across the *cpmp* gene and narrow down the genetic region of interest which will constitute your marker, you need to visualize the allele profile across this gene for the multiple samples. This allows you to identify which regions are more mutated than others. For this purpose, you will need to plot the mutation frequencies provided in the .txt file and identify a suitable region to target (according to criteria above). The R script "*exercise1_TES_notebook.R*" contains the steps to guide you through the data processing and finding this region. Address the questions and run the commands from the script to obtain your target region, then include here a plot of the resulting mutation frequencies. Can you already identify some polymorphic regions which can be good candidate for markers?

❖ **Plot of the mutation frequencies:**



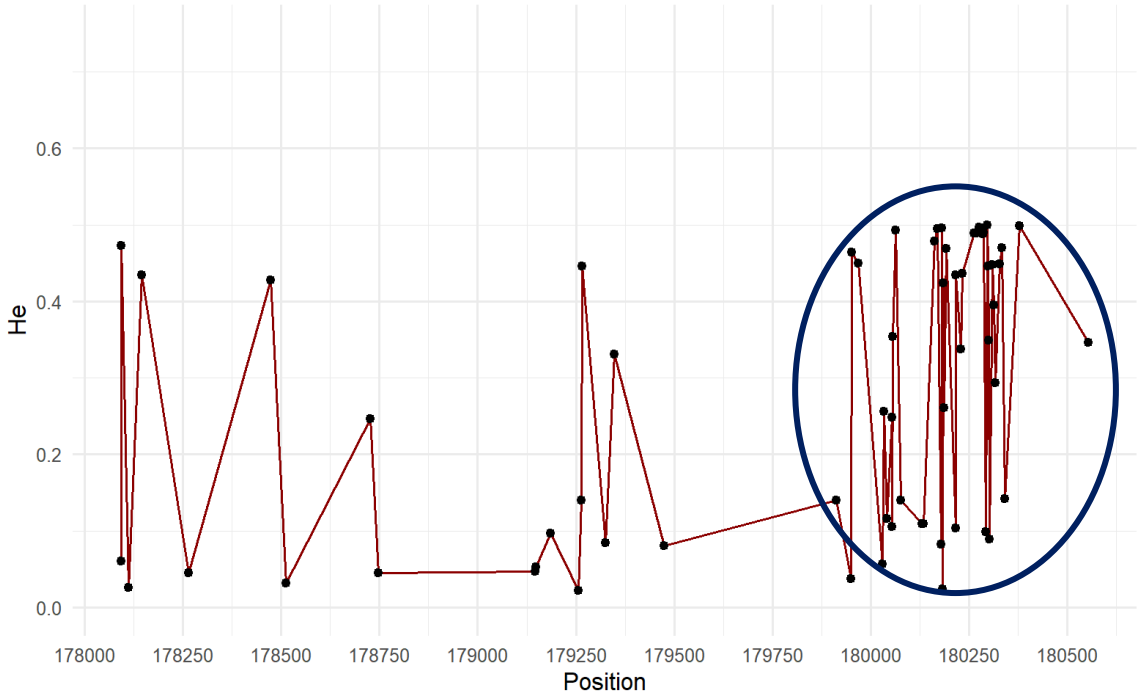
❖ **Comment:**

The mutation frequency plot shows the distribution of observed genetic variations across different positions of the cpmp gene. Most positions exhibit low mutation frequencies, but there **is a notable increase in variation around positions 180,000 and beyond**.

Since a good marker should capture high genetic diversity within a limited region (≤ 300 nucleotides), this highly mutated segment around position 180,000 could be a strong candidate for a targeted marker region.

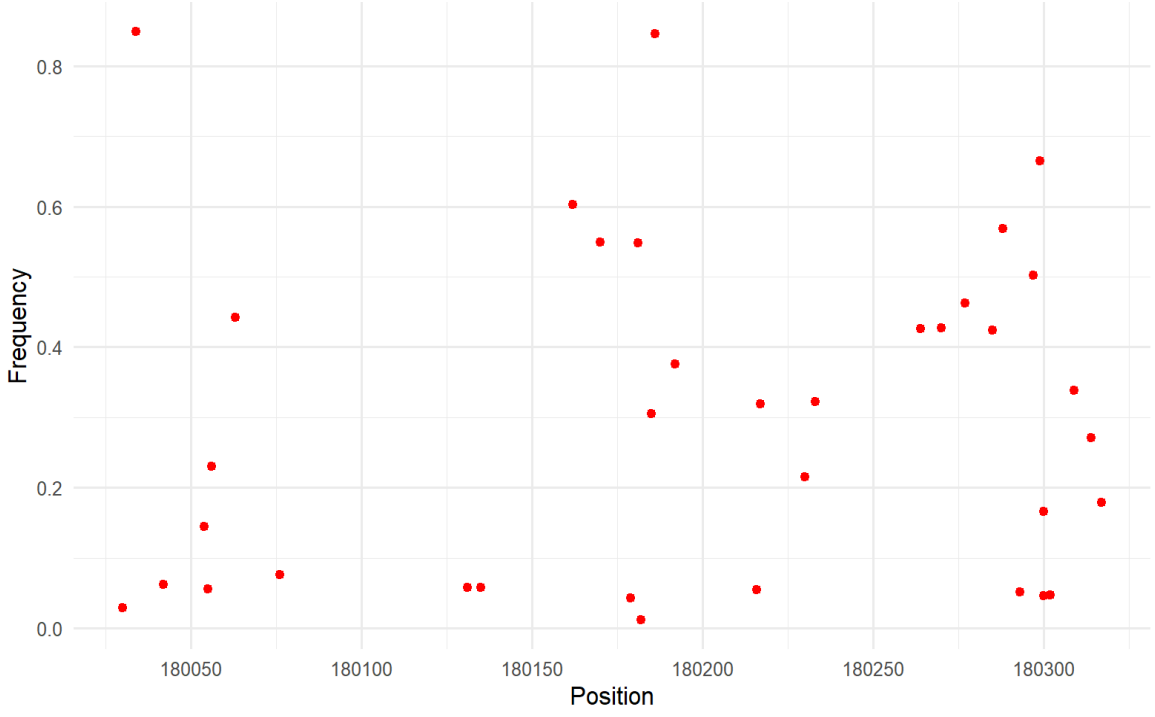
3. Calculate the expected heterozygosity at each position displaying genomic variation in the *cpmp* gene. Include a plot which contains the expected heterozygosity information. Highlight a region of 300 nucleotides that you would choose as a marker for the therapeutic efficacy study.

❖ **Plot of heterozygosity**



❖ **Comment:**

As expected from the previous plot, the zone from 180.000 Kb shows a higher heterozygosity than other zones. Accordingly, we decided to choose the position from 180025 Kb to 180325 Kb as the region of interest. Focusing on this area, we can replot the frequencies of mutations, and we get this new plot, which shows the genomic variations.



Part II & III: Haplotype calling, classification of infections and treatment efficacy estimation

Following a similar approach as shown in part I, your team has identified several other target regions (markers) to use for identifying the different parasite strains within malaria infections. These markers are highly polymorphic regions within the genes: *cpmp, ama-1-D2, and msp7*. Your colleagues then collected several patient samples on Day 0 and Day X and ran the AmpSeq experiments for the selected markers. In these experiments, the selected targeted regions were amplified in each pair of infections (Day 0 and Day X) and then sequenced.

Sequence alignment of samples

For this part of the exercise, you will need to copy folder Case study part 2/ from the course Exercises/ folder to your Home directory on your computer.

1. Look at the extensions (.fna, fastq.gz, etc.) of the files in the folder that you have just copied. Can you already identify what each file corresponds to (reference sequence, sequenced patient samples)? Which patient sequencing file corresponds to Day 0 and which one to Day X?

❖ **Reference Genome files:**
- *reference_HB37.fna :* which contains the reference genome.
- *reference_HB37.fna.fai:* is the index file with chromosome locations and mutations.

❖ **Patient sample files (FASTQ format):**
- *A00194_BC_Fw_2-Rv_2_cpmp_F.fastq:* patient sequencing file corresponding to ***Day 0***
- *A00221_BC_Fw_6-Rv_2_cpmp_F.fastq:* patient sequencing file corresponding to ***Day X***

2. Open the R script *exercise2_TES.R* and follow the R commands in order to align your sample pair to the *cpmp* gene. Investigate the files created in the process and mention what they correspond to. Which are the files containing the aligned reads? How many reads could be aligned?

❖ **Files Containing Aligned Reads**

The aligned reads are stored in **BAM files**, while the corresponding **BAI files** serve as index files for efficient access to the alignments.

✓ **Day 0 sample:**
- *A00194_BC_Fw_2-Rv_2_cpmp_F_2d386e2e93b.bam:* Contains the sequence alignment of the patient's *day 0* sequencing data to the reference genome.
- *A00194_BC_Fw_2-Rv_2_cpmp_F_2d386e2e93b.bam.bai:* Index file for the corresponding BAM file.

✓ **Day X sample:**
- *A00221_BC_Fw_6-Rv_2_cpmp_F_2d3839a94750.bam:* Stores the sequence alignment of the patient's *day X* sequencing data to the reference genome.
- *A00221_BC_Fw_6-Rv_2_cpmp_F_2d3839a94750.bam.bai* - Index file for the corresponding BAM file.

❖ **How many reads could be aligned:**

Genome	Seqlength	Mapped	Unmapped
Sample_D0	4142180	15916	11435
Sample_DX	4142180	1463	29463

Table 1: The sequence reads analysis results

✓ **Comment:**
For *Day 0, only* 15,916 reads were successfully mapped, while 11,435 reads remained unmapped. For *day X, o*nly 1,463 reads were successfully mapped, whereas 29,463 reads remained unmapped, indicating a lower alignment rate compared *to day 0.* Hence, there is a significant drop in mapped reads from *day 0* to *day X*.

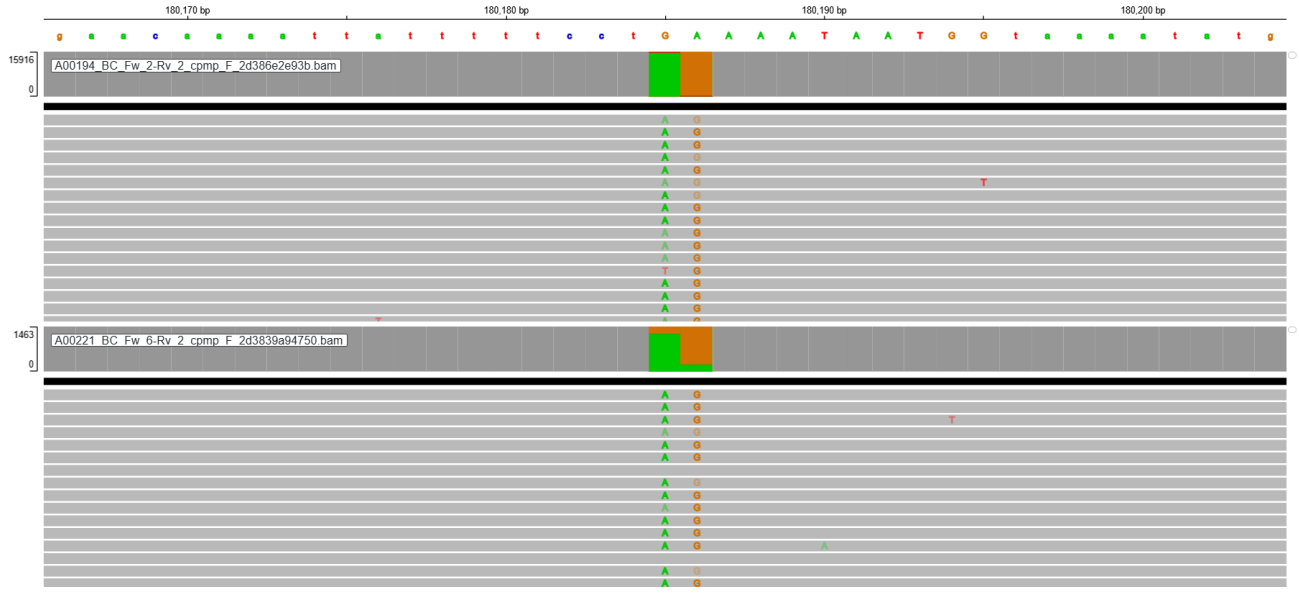
3. To see how the previously chosen target region for the *cpmp* gene is varying for a patient sample at day 0 and at day X, you can visualize the alignment results using a web browser. You will need to access the online genome browser IGV (https://igv.org/app/).
First you will need to load the reference sequence (.fna extension) together with its index (.fai extension) in the browser (menu Genome -> Local file …), then to add each aligned sample (.bam extension) together with its index (.bai extension) as a track (menu Tracks -> Local file). Afterwards, you can navigate to the region of interest that you have identified at Part I and visualize the aligned reads from the samples at Day 0 and Day X. Overall, do you observe any sequence variation for the targeted region you have identified yesterday? Can you identify some shared haplotypes between Day 0 and Day X? You can include a screenshot of the alignment from the browser as well as haplotypes you can identify.

❖ **Plot of the alignment results highlighting different types of sequence variations**

✓ **Comment:**

There are observable sequence variations region when comparing the *day 0* to *day X* samples. Different nucleotide positions can be seen between the two samples. We can also observe that the number of mapped reads in *day X* is significantly lower than on *day 0*.

❖ **Plot highlighting the identifed haplotype**



✓ **Comment:**

We have regions showing conserved sequences between *day 0* to *day X*, as there is consistency in position *G* and *A* bases, suggesting we have a haplotype. While some mutations have occurred, certain nucleotide sequences remain unchanged (in black) between the two days, indicating retained haplotype segments.

Haplotype identification with the DADA2 R package

DADA2 is a commonly used package for haplotype identification. It works on paired sequencing reads. The package has a dedicated page with tutorials and resources at https://benjjneb.github.io/dada2/tutorial.html.

1. Follow the commands from the Case study part 3/ *exercise3_TES_part1.R* to identify the haplotypes from your sample at Day 0 and Day X. Paste here the identified haplotype sequences and highlight the mutations.

❖ **Following the identified haplotypes with the differences in mutations in red:**

➢ **Haplotype 1**

```
[1] "T" "A" "T" "A" "T" "A" "T" "G" "G" "A" "T" "T" "G" "T" "T" "A" "T" "A" "A" "T" "G" "A"
"A" "A" "C" "G" "A" "G" "A"
[30] "T" "A" "A" "T" "A" "A" "T" "A" "C" "A" "T" "T" "T" "T" "T" "A" "A" "A" "T" "A" "G" "A"
"A" "T" "A" "A" "G" "C" "C"
[59] "A" "T" "G" "A" "A" "G" "G" "A" "A" "T" "C" "A" "A" "A" "A" "A" "A" "T" "C" "G" "T" "T"
"A" "T" "T" "A" "A" "T" "A"
[88] "A" "C" "A" "G" "G" "T" "A" "T" "T" "A" "T" "A" "A" "A" "C" "G" "A" "A" "A" "A" "T" "A"
"T" "T" "A" "A" "A" "T" "A"
[117] "T" "G" "A" "A" "A" "A" "T" "A" "T" "G" "G" "A" "A" "G" "C" "T" "A" "T" "A" "G" "G"
"T" "A" "T" "C" "A" "G" "A" "T"
[146] "C" "C" "T" "A" "T" "T" "T" "A" "A" "C" "T" "A" "C" "G" "T" "T" "G" "A" "A" "C" "A"
"A" "A" "A" "T" "T" "A" "T" "T"
[175] "T" "T" "T" "C" "C" "T" "G" "A" "A" "A" "A" "T" "A" "A" "T" "G" "G" "T" "A" "A" "A"
"A" "T" "A" "T" "G" "T" "T" "T"
[204] "T" "A" "A" "T" "G" "A" "T" "A" "A" "C" "A" "T" "A" "T" "G" "T" "G" "A" "A" "A" "A"
"G" "G" "A" "G" "A" "A" "T" "A"
[233] "A" "T" "A" "C" "A" "T" "A" "T" "A" "A" "T" "A" "T" "T" "A" "C" "T" "G" "A" "G" "A"
"A" "T" "T" "C" "A" "A" "C" "A"
[262] "A" "C" "A" "A" "A" "C" "G" "A" "T" "C" "A" "G" "A" "A" "A" "T" "A" "T" "A" "G" "A"
"A" "C" "T" "A" "T" "A" "C" "G"
[291] "T" "G" "G" "T" "C" "G" "A" "A" "T" "A" "A" "G" "T" "G" "C" "A" "A" "A" "T" "C" "T"
"T" "G" "G" "A" "A" "A" "C" "G"
[320] "A" "T" "T" "T" "G" "G" "A" "T" "G" "C" "T" "A" "T" "T" "G" "T" "T" "C" "A" "A" "C"
"T" "C" "T" "T" "T" "G" "T" "T"
[349] "T" "A" "T" "A" "A" "A" "G" "C" "A" "C" "G" "T" "A" "T" "T" "C" "T" "A" "A" "A" "A"
"G" "A" "A" "A" "A" "T" "A" "T"
[378] "A" "T" "T" "C"
```

- **Haplotype 2**

```
[1] "T" "A" "T" "A" "T" "A" "T" "G" "G" "A" "T" "T" "G" "T" "T" "A" "T" "A" "A" "T" "G" "A"
"A" "A" "C" "G" "A" "G" "A"
[30] "A" "A" "A" "T" "A" "A" "T" "A" "C" "A" "T" "T" "T" "T" "T" "A" "A" "A" "T" "A" "G" "A"
"A" "T" "A" "A" "G" "A" "A"
[59] "C" "T" "G" "A" "A" "C" "A" "G" "T" "C" "A" "A" "A" "A" "A" "A" "T" "C" "G" "T" "T"
"A" "T" "T" "A" "A" "T" "A"
[88] "A" "C" "A" "G" "G" "T" "A" "T" "T" "A" "T" "A" "A" "A" "C" "G" "A" "A" "A" "A" "T" "A"
"T" "T" "A" "A" "A" "T" "A"
[117] "T" "G" "A" "A" "A" "A" "T" "A" "T" "G" "G" "A" "A" "G" "C" "T" "A" "T" "A" "G" "G"
"T" "A" "T" "C" "A" "G" "A" "T"
[146] "C" "C" "T" "A" "T" "T" "T" "A" "A" "C" "T" "A" "C" "G" "T" "T" "G" "A" "A" "C" "A"
"A" "A" "A" "T" "T" "A" "T" "T"
[175] "T" "T" "T" "C" "C" "T" "A" "G" "A" "A" "A" "T" "A" "A" "T" "G" "G" "T" "A" "A" "A"
"A" "T" "A" "T" "G" "T" "T" "T"
[204] "T" "A" "A" "T" "G" "A" "T" "A" "T" "C" "A" "T" "A" "T" "G" "T" "G" "A" "A" "A" "A"
"G" "G" "A" "G" "A" "A" "T" "A"
[233] "A" "T" "A" "C" "A" "T" "A" "T" "A" "A" "T" "A" "T" "T" "A" "C" "T" "G" "A" "G" "A"
"A" "T" "T" "C" "A" "A" "A" "A"
[262] "A" "C" "A" "A" "C" "C" "G" "A" "T" "C" "A" "A" "A" "A" "A" "T" "A" "T" "A" "T" "A"
"A" "C" "T" "A" "T" "A" "C" "G"
[291] "T" "G" "A" "T" "G" "G" "A" "A" "T" "G" "G" "A" "T" "G" "A" "A" "C" "A" "T" "C" "T"
"T" "A" "G" "A" "A" "A" "C" "G"
[320] "A" "T" "T" "T" "G" "G" "A" "T" "G" "C" "T" "A" "T" "T" "G" "T" "T" "C" "A" "A" "C"
"T" "C" "T" "T" "T" "G" "T" "T"
[349] "T" "A" "T" "A" "A" "A" "G" "C" "A" "C" "G" "T" "A" "T" "T" "C" "T" "A" "A" "A" "A"
"G" "A" "A" "A" "A" "T" "A" "T"
[378] "A" "T" "T" "C"
```

- **Haplotype 3**

```
[1] "T" "A" "T" "A" "T" "A" "T" "G" "G" "A" "T" "T" "G" "T" "T" "A" "T" "A" "A" "T" "G" "A"
"A" "A" "C" "G" "A" "G" "A"
[30] "T" "A" "A" "T" "A" "A" "T" "A" "C" "A" "T" "T" "T" "T" "T" "A" "A" "A" "T" "A" "G" "A"
"A" "T" "A" "A" "G" "C" "C"
[59] "A" "T" "G" "A" "A" "G" "G" "A" "A" "T" "C" "A" "A" "A" "A" "A" "A" "T" "C" "G" "T" "T"
"A" "T" "T" "A" "A" "T" "A"
[88] "A" "C" "A" "G" "G" "T" "A" "T" "T" "A" "T" "A" "A" "A" "C" "G" "A" "A" "A" "A" "T" "A"
"T" "T" "A" "A" "A" "T" "A"
[117] "T" "G" "A" "A" "A" "A" "T" "A" "T" "G" "G" "A" "A" "G" "C" "T" "A" "T" "A" "G" "G"
"T" "A" "T" "C" "A" "G" "A" "T"
[146] "C" "C" "T" "A" "T" "T" "T" "A" "A" "A" "A" "G" "A" "G" "T" "T" "G" "A" "A" "C" "C"
"A" "A" "A" "G" "G" "A" "T" "T"
[175] "T" "T" "C" "C" "C" "T" "G" "G" "A" "A" "A" "T" "A" "T" "T" "G" "G" "T" "A" "A" "A"
"A" "T" "A" "T" "G" "T" "T" "T"
[204] "T" "A" "A" "T" "G" "A" "T" "A" "T" "C" "A" "T" "A" "T" "G" "T" "G" "A" "A" "A" "A"
"G" "G" "A" "G" "A" "A" "T" "A"
[233] "A" "T" "A" "C" "A" "T" "A" "T" "A" "A" "T" "A" "T" "T" "A" "C" "T" "G" "A" "G" "A"
"A" "T" "T" "C" "A" "A" "A" "A"
[262] "A" "C" "A" "A" "C" "C" "G" "A" "T" "C" "A" "A" "A" "A" "A" "T" "A" "T" "A" "T" "A"
"A" "C" "T" "A" "T" "A" "C" "G"
[291] "T" "G" "A" "T" "G" "G" "A" "A" "T" "G" "G" "A" "T" "G" "A" "A" "C" "A" "T" "C" "T"
"T" "A" "G" "A" "A" "A" "C" "G"
[320] "A" "T" "T" "T" "G" "G" "A" "T" "G" "C" "T" "A" "T" "T" "G" "T" "T" "C" "A" "A" "C"
"T" "C" "T" "T" "T" "G" "T" "T"
[349] "T" "A" "T" "A" "A" "A" "G" "C" "A" "C" "G" "T" "A" "T" "T" "C" "T" "A" "A" "A" "A"
"G" "A" "A" "A" "A" "T" "A" "T"
[378] "A" "T" "T" "C"
```

2. What is the multiplicity of infection (MOI) of the sample?

❖ **MOI of the Sample:**
The multiplicity of infection of the sample represents the number of different parasite strains coexisting in the sample. Here, it corresponds to the number of distinct haplotypes found in the sample. Then, since, we have three different haplotypes, we can conclude that the **MOI is 3.**

Classification of recrudescence and new infection

Using your established methodology from the previous parts, the malaria surveillance team in Weveland has performed the complete bioinformatics analysis and has identified for each sample pair of several patients (day 0 and day X for each patient) the different haplotypes for four different markers: ama1, cpmp, cpp and msp7. You will apply the 2/3 and 3/3 algorithms to identify the recrudescences and new infections within the patient samples, and then calculate the treatment efficacy.

The results with the identified haplotypes for each marker and samples are provided in the file haplotypes.csv.

1. Continue following the provided R script (*exercise3_TES_part2_notebook.R).* Load the table with the identified haplotypes (haplotypes_in_samples.csv). What are the different columns in the table? How many samples were analyzed?

❖ **Columns in the table**
- **X:** Data generated index.
- **SampleName:** Sample IDs for each patient as per the markers **for both day 0 and day X**
- **MarkerID:** The genetic marker used for classification from the sample obtained: cpmp, ama1-D3, and msp7 for both day 0 and day X.
- **Haplotype:** The haplotype observed in the sample for the patient's sample for both day 0 and day X.
- **Reads:** The number of sequencing reads for the observed haplotype in the sample.
- **MOI:** The estimated MOI value for the sample.

❖ **Samples were analyzed**

Samples from 20 patients were analyzed at both *day 0* and *day X*.

2. There are two commonly used algorithms to differentiate recrudescence from new infection: 2/3 (2 out of 3 markers have to show a recrudescence) and the 3/3 algorithm (3 out of 3 markers have to show recrudescence). Calculate the treatment efficacy using the two classification results (2/3 and 3/3). Do you observe any differences?

❖ **Treatment efficacy formula**

$$T = \left(1 - \frac{Number\ of\ recrudescences}{Total\ number\ of\ patient\ with\ Adequate\ follow-up}\right) \times 100$$

- **For the 3/3 algorithm:**

$$T = 45\%$$

- **For the 2/3 algorithm:**

$$T = 20\%$$

❖ **Comparison:**
The treatment efficacy differs depending on the classification algorithm. The 3/3 algorithm suggests that the current drug regimen may be failing in nearly half of the population since it is more sensitive to detecting **true recrudescence.** On the other hand, the 2/3 algorithm **allows some variation** as it may also capture reinfections rather than pure resistance; hence, a 20% rate still raises concern about drug failure.

3. Based on the results from the previous question and your expertise, what would be your recommendations for the National Malaria Control Program of Weveland with regard to the antimalarial therapy used in the country? Should the therapy be changed?

❖ **Recommendations:**
From the above results, both algorithms, the 3/3 and 2/3, indicate a potential decline in the efficacy of the current antimalarial treatment in Waveland. Looking to that, we can suggest the following actions:

- **Conduct further surveillance:** expand the study with larger sample sizes across different regions to confirm the observed trends.
- **Assess potential drug resistance:** by investigating molecular markers of resistance in persistent infections and analyzing drug pressure and parasite clearance rates in treated patients.
- **Consider alternative treatment strategies:** if further studies confirm low efficacy, explore the introduction of alternative or combination therapies.
- **Strengthen infection classification methods:** use genomic data and MOI analysis to distinguish recrudescence from reinfection more accurately and integrate whole-genome sequencing where possible to complement haplotype-based classification.

4. After learning about targeted amplicon sequencing and its use in therapeutic efficacy studies, what do you see as the most sigtnificant challenges to its implementation in malaria-endemic countries? Reflect on issues such as panel design, bioinformatics pipelines, and infection classification, and explain why you consider them challenging.

❖ **Challenges with the implementation of amplicon sequencing in malaria-endemic countries**

➢ **Panel Design**
- **Challenge:** Selecting the most relevant single-nucleotide polymorphisms (SNPs) and drug-resistance markers requires extensive research, along with large-scale genomic data collection across diverse malaria-endemic regions.
- **Why:** The complex life cycle of the Plasmodium parasite involves multiple stages (mosquito, liver, and blood stages), each with distinct selective pressures, contributing to high genetic diversity. This makes it difficult to design a single, universally effective SNP panel. Additionally, primer efficiency may vary across different Plasmodium species and strains, potentially leading to biased detection of resistance markers.

➢ **Bioinformatics pipelines:**

- **Challenge:** Processing large sequencing datasets requires advanced computational infrastructure and standardized analysis workflows.
- **Why:** Many malaria-endemic regions lack access to high-performance computing resources, making large-scale sequencing analysis difficult. Additionally, there is no globally standardized bioinformatics pipeline for AmpSeq in malaria surveillance. For example, the WHO has not recommended a single standard algorithm for distinguishing recrudescence from reinfection, leading to inconsistencies across studies. Some studies use allelic frequency-based methods, while others rely on haplotype-matching approaches, resulting in potential discrepancies in drug efficacy estimates.

➢ **Infection classification:**

- **Challenge:** Distinguishing between recrudescence (drug failure) and reinfection (new infection) is difficult due to genetic recombination and mixed infection.
- **Why:** In high-transmission areas, patients are frequently exposed to multiple Plasmodium strains, making it hard to determine if a persistent infection is due to drug resistance or reinfection with a similar strain. This affects the accuracy of treatment efficacy studies and could lead to misclassification of treatment outcomes.