Case study:

Determining the malaria treatment efficacy in Weveland

Détagnon Bernic GBAGUIDI

Henriette DUKUZIMANA

**Please read carefully and respond to the questions below by writing short sentences/paragraphs below each question and including plots from the coding exercises. This assignment together with the associated R scripts are due on the 21st of April, and should be sent for grading to** [**monica.golumbeanu@swisstph.ch**](mailto: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 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 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.

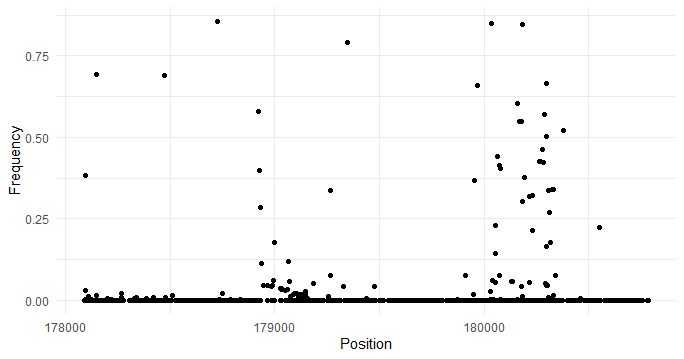
**Questions:**

For this part of the exercise, you will need to access the google drive link for the course:

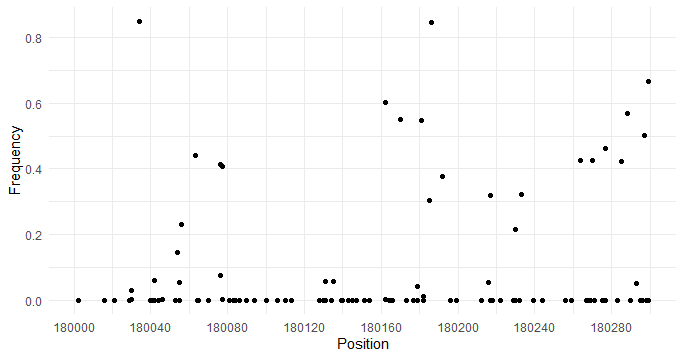
<https://drive.google.com/drive/folders/1fdFFq6WIzSE1dmcMLcSgAXWzI_twr3TE?usp=sharing>, then copy folder Day1/ to a folder in your Home directory on your computer.

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?

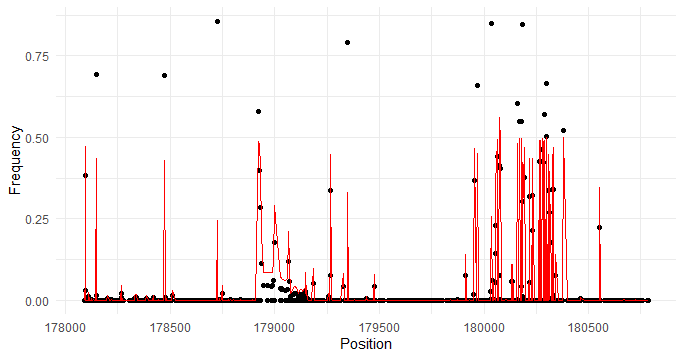
Answer: cpmp stand for conservative membrane plasmodium protein. It one of the gene use as marker in microbiology surveillance. Used as marker this should be highly polymorphic.Also as we can see in the following plot it has some region susceptible for frequent mutations.



1. 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 mutation 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\_parasite\_genomics.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 selected region.



1. BONUS: Calculate the expected heterozygosity at each position displaying genomic variation in the cpmp gene. Include a plot which contains the expected heterozygosity information. What do you observe for the region that you have previously selected?



We can see that the previous region selected (180000-180300) have high variation of the expected heterozygosity.

**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 access the google drive link for the course:

<https://drive.google.com/drive/folders/1fdFFq6WIzSE1dmcMLcSgAXWzI_twr3TE?usp=sharing>, then copy folder Day2/ to a folder in 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, samples)? Which sample file corresponds to Day 0 and which one to Day X?

**Answer:**

The file with extension .fna correspond to the reference sequence.

The file with extension .fastq.gz correspond to the samples.

A00194\_BC\_Fw\_2-Rv\_2\_cpmp\_F.fastq.gz corresponds to day 0 and

A00221\_BC\_Fw\_6-Rv\_2\_cpmp\_F.fastq.gz corresponds to day X.

Furthermore file with extension .fnai contains index to facilitate rapid retrieval of sequence information and also .bam and .bai correspond respectively to alignment of

sequencing reads to a reference genome and the corresponding index file.

1. Open the R script *exercise2\_parasite\_gentics.R* and follow the R commands in order to align your sample pair to the cpmp region. 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?

Answer:

We have two main kind of files that has been created. Firstly file with extension .bam and .bai which correspond to the alignment of sequencing reads to a reference genome for each sample.

Secondly we have qc\_report.pdf that report some statistics about the alignment.

The following filles has been created:

* A00194\_BC\_Fw\_2-Rv\_2\_cpmp\_F\_aec301b1a7a.bam ( containing alignement for day 0)
* A00194\_BC\_Fw\_2-Rv\_2\_cpmp\_F\_aec301b1a7a.bam.bai (containing index of alignment for day 0)
* A00194\_BC\_Fw\_2-Rv\_2\_cpmp\_F\_aec301b1a7a.bam.txt (txt.file containing some informations )
* A00221\_BC\_Fw\_6-Rv\_2\_cpmp\_F\_aecbbd1f03.bam ( containing alignement for day X)
* A00221\_BC\_Fw\_6-Rv\_2\_cpmp\_F\_aecbbd1f03.bam.bai (containing index of alignment for day X)
* A00221\_BC\_Fw\_6-Rv\_2\_cpmp\_F\_aecbbd1f03.bam.bai.txt (txt.file containing some informations )
* QuasR\_log\_aec38f172b9 (log file )

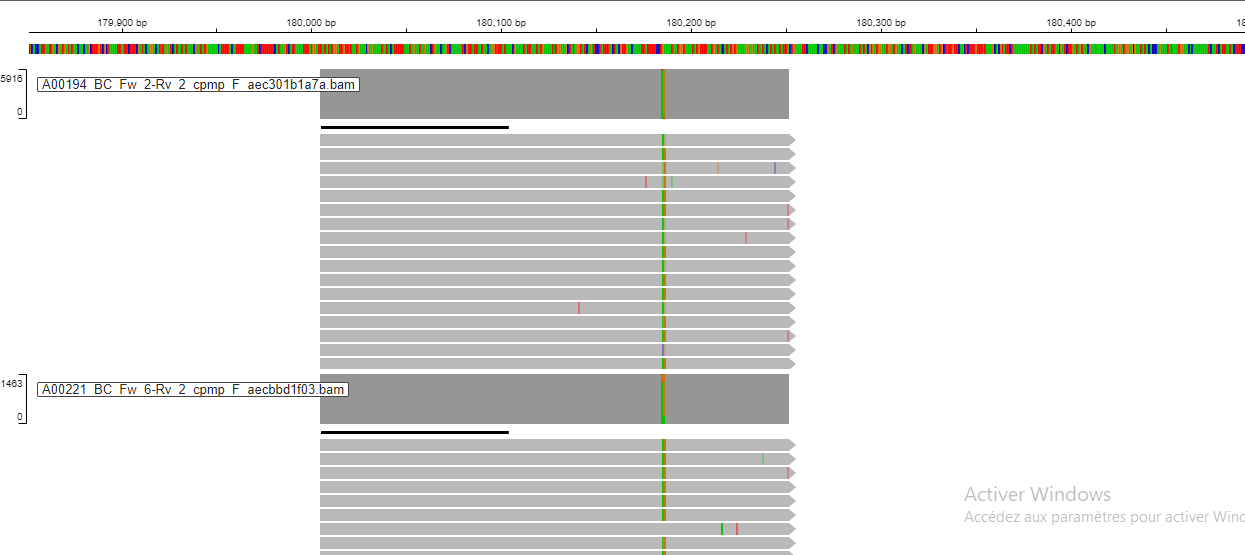
For the sample of Day 0 we have 15916 (58.2%) reads aligned while for Day X we have 1463 (4.7%).

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 can use 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.

1. 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.

Answer:

The region targeted is between 180000 and 180400. We can notice that there some mutation in some reads but mainly as the screenshot below is showing , we can identify one shared haplotype between Day 0 and Day X.



**Classification of recrudescence and new infection**

The malaria surveillance team in Weveland has performed the complete bioinformatics analysis and has identified for each sample pair (day 0 and day X for each patient) the different haplotypes for the different markers. 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 (*exercise2\_parasite\_genetics.R).* Load the table with the identified haplotypes (haplotypes\_in\_samples.csv). What are the different columns in the table?

**Answer:**

The different columns of the table is : "SampleName" "MarkerID" "Haplotype" and "Reads"

1. 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?

**Answer:**

The treatment efficacy of the 2/3 algorithm is :0.2

The treatment efficacy of the 3/3 algorithm is : 0.25

(details of how this value is obtained are in the .R file)

1. Write a paragraph where you present your opinion about the challenges of using targeted amplicon sequencing estimating treatment efficacy in therapeutic efficacy studies in malaria endemic countries. Address topics such as panel design, bioinformatics analysis and classification algorithms.

Answer:

Targeted amplicon sequencing holds promise in assessing treatment efficacy in malaria-endemic regions, yet it navigates a labyrinth of challenges. Panel design is pivotal, demanding meticulous curation to encompass genetic diversity and pertinent drug resistance markers. Bioinformatics analysis encounters hurdles in data processing, especially with high-throughput sequencing generating copious datasets. Moreover, classification algorithms such as 2/3 and 3/3 proposed by the World Health Organization (WHO) represent commendable efforts to standardize the evaluation of treatment efficacy, offering clear guidelines for categorizing treatment outcomes based on parasitological and clinical parameters. However, while they provide a structured framework, their application may encounter limitations in capturing the complexity of treatment responses, particularly in regions with diverse parasite populations and varying drug susceptibilities. The reliance on predefined thresholds may overlook subtle changes in parasite clearance dynamics or emerging resistance patterns. Hence, the confluence of challenges in targeted amplicon sequencing underscores the need for interdisciplinary collaboration, innovative methodologies, and a nuanced understanding of local malaria dynamics to harness its full potential in evaluating treatment efficacy.