steps to SNPs

- fastp.pl
  - By the use of fastp, reads are trimmed on quality. Every base in a read has a score for the chance of incorrect calling. This score is known as Phred.
- minimap2
  - By the use of `minimap2`, reads are aligned to the reference genome (Rattus Norvegicus) and these alignments are saved as BAM files (4 per organism; for each organism 4 runs are sequenced) all files are sequenced in two directions(paired end), but these are used and combined in this step.
- before calling
  - Combine the 4 BAM files for each organism to 1 BAM file, so there is 1 'huge' alignment of all reads per sample. Changes to the sample names of all reads will apply to the relevant samples in each file, otherwise, HaplotypeCaller will fail to call. Sort the BAM files to make it easier to find reads mapped at a certain point in the genome.
- Calling
  - Variant Calling refers to finding variants on the alignments.
  - Calling of variants is first tried with 'HaplotypeCaller'. The Cambodia sample has been successfully called with this method. HaplotypeCaller could be runned in a parallel way by using a Scala script and Queue, which will split the whole BAM file up into specific genomic ranges, so multiple HaplotypeCaller processes could process proceed with different genomic ranges alongside each other.
  - Calling is then done using `bcftools`. It is only used to find SNPs, thus insertions/deletions of bases are not taken into account. The result of the calling was a bcf file (binary variant calling format). This is done for all samples, including the Cambodia sample. bcftools is more complicated to parallelize, except for a small part of the algorithm.
- before database
  - The 8 processed bcf files are imported into a database. This database has columns based on the bcf/vcf file with the following modifications:
    - The id(3rd) column and the filter(7th) column are not in the database. These are empty for all records (or at least about 3 samples).
    - The format column is the same ('GT:PL') for all records and is not in the database.
    - The info column is replaced with the depth of the reads.
    - Genotype and phred score for each genotype are two separate columns.
    - Following these columns are two columns describing the distance between the previous and next SNP. If this is not available, the values are rendered -1 (thus filtered out in some next step currently.)
    - At last there is a column containing a number referencing to the sample where it belongs to. The table that contains information on which numbers are mapped to which sample is called sample-enum.csv (on this github repository)
  - There is also a database created with only unique SNPs and the Genotype/Phred columns per sample.
- database
  - It is filtered on multiple ways:
    - Is it homozygote, i.e., is the SNP fully shared over all 8 organisms? If so, the SNP will not detect much.
    - Is the general phred quality score higher than 99?
    - Is the number of reads on the position of the SNP (*coverage*) greater than 16 and smaller than 110?
    - Is there no other SNP 300 bases upstream or downstream?
  - All those conditions must be true to keep the SNP.
  - The chromosome/position pairs that belong to genuine SNPs according to the filter steps are saved as a file, and later as a table in the database.
  - These SNPs are then fully exported to a file on disk.
- Consensus genome
  - Consensus genomes of the rats are made by calling all types of variants for a BAM file, and use these variants.
- BLASTing
  - The sequence of 250 bases before and after the SNP is collected from the reference genome(*R. Norvegicus*), and BLASTed against all the constructed consensus genomes of the 8 rats. The SNPs belonging to sequences which are matched more than one, or zero times, are rejected.
- Final Selection
  - more diverse SNPs are prefered, SNPs that only apply to one organism are avoided as much as possible.

Position (a), alelle (Adenine), rat (North New Zealand)

Position (b), alelle (Thymine), rat (North New Zealand)

Position (a), alelle (Adenine), rat (Laos (1st))

Position (c), alelle (Adenine), rat (North New Zealand)

Position (a), alelle (Adenine), rat (Thailand Tak province)

Position (a), alelle (Adenine), rat (Loas (2nd))

Position (d), alelle (Adenine), rat (Prachuap Kiri Khan)

...

...

4 of them has Adenine at position a;
If the other 4 has Adenine as well, filter this out (don't let this pass the filter) for being homozygote

Is the (phred) quality lower than 100:
Filter this out

Is the coverage of the SNP **not** between 16 and 110:
Filter the SNP out

Is there another SNP within 300 bases upstream or downstream: Filter the SNP out

Low coverage (~11)

Higher coverage (~20)