**minion\_Genotyper Manual**

**Folder content**

minion\_Genotyper.py: This is the main program

kmerDB folder: This is the folder where the kmers that are used for the

genotyping are kept

depositedSequences\_codes.txt This is a file were the genotyping codes for all the 244

deposited genomes are reported

example\_kmersCount.txt This an example of how the kmer count will be reported for

each read

example\_readsCodes.txt This is an example of how the reads are coded

example\_statistics.txt This is an example of the main output file (see below)

codeTable.xls This is the table that links the used one letter code with the original genotype codes.

**Running the software**

The software runs with the following command within the minion\_Genotyper folder:

*python minion\_Genotyper.py inputFileName minimumReadLength outPrefix*

inputFileName: fastq file name with its complete path

minimumReadLength; Minimum length for a read to be included in the downstream

analyses (so far 50000 proved to be a good number)

outPrefix: this is a name that will constitute the prefix of all the output files (e.g.

the outPrefix example will generate the output files that are now

present in the folder

**Output file**

The file \_statistics.txt is the main output folder.

In the first section you will find for each hypervariable gene the code of the genotype that has been found and how many times this occurred.

In the second section you will find the reconstructed strains that are reported as the genotypes string code. If the string code has been previously observed in any of the 244 deposited HCMV genomes, then the name of such genome will be reported close to the coded strain. If not, a “No match found” will be reported.

In the third section some “reliability” statistics are reported. Each found strain code is divided in trimers and all the code is scanned by using a sliding window (window size = 3, window step = 1). The software will count how many times that trimer has been called in the reads. Strain codes featuring some of the trimers with low count as compared to the other observed values are probably artefact and possibly need to be removed. As a way of example, in the file example\_statistics.txt all strain codes have at least one trimer that is present twice or less in the reads, with the exception of the last two (which in this case are also the only ones to get a match in the 244 deposited genomes as seen in the section before). The number of unreliable strain codes should be very low and it is not in this case just because I used a very low number of reads to test the software. You can use this section to copy and paste the trimer counts for each strain code into Excel and calculate for each trimer position the frequency and then average all of them to obtain the frequency of each strain in a mixed strain infection.

The sensitivity of the software can be tuned by changing the following line in the main software

detectionTresholdPercentage = 0.002

If you low down that 0.002 the sensitivity will increase, but of course in this case you may risk to take in some low frequent / possibly artifactual strains.