estMOI: Estimating Multiplicity of Infection using parasite deep sequencing data
(version 1.03)

1. For help and available options

```
estMOI_1.03 -h
```

2. USAGE:

```
estMOI_1.03 <sample.bam>  <vcf.gz>  <ref.fasta> [OPTIONS]
```

OPTIONS:

```
 --readl=int       Minimum read length [default 76 ]
 --maxsnp=int      Maximum number of SNPs [default 10 ]
 --mincov=int      Mininum number of reads over a SNP [default 10]
 --minhap=int      Minimum frequency of haplotypes to consider[def. =3]
 --maxfact=int     Percentile cutoff for adjusting genomewide MOI
                   estimate [default 90]
 --mindis=int      Minimum dist. between any two SNPs [def.=10]
 --maxdis=int      Maximum dist.between the first and last SNP[def.500]
 --flank=int       Flanking size for excluded regions [def. 500]
 --out=string      Output file prefix [default estMoi]
 --tmp             Do not delete temporary working directory
 --debug           Debug output [default 0]
```

3. OUTPUT

Two files are generated for every run of estMOI.
The first file (*.log) contains haplotypes and MOI estimates for all combinations of three SNPs that fulfill the distance constraints. An overall estimate of MOI per chromosome is also shown. An example LOG file is shown below

```
SampleID    Pf3D7_12_v3     2572 2626 3031  2
        # Sample01 Pf3D7_12_v3 2572 2626 3031 Hapotype:      T G T  22
        # Sample01 Pf3D7_12_v3 2572 2626 3031 Hapotype:      T G A  3
```

The first line is used for further analysis (using the script reRun_estMOI). The columns represent:

sampleID, Chromosome, SNP1, SNP2, SNP3 and NumberOfHaplotypes

The remaining lines show the genotype calls of the haplotypes and the frequency of each haplotype. Eg. The haplotype TGA is seen 22 times while TGA is only seen three times. Note that a minimum haplotype frequency of 3 (i.e the default) is used in the example above. However, if we were to re-run by increasing the minhap to 4, there will only be one haplotype in the region of interest.

The second file (*.txt) contains a summary of final MOI estimates from a genomewide computation of haplotype counts. For the example data set, the output file has 623 regions with a single MOI estimates; 22 regions with a MOI of 2 and 15 regions with a MOI of three (columns 1 and 2 of the output file respectively; see below).

```
#MOI    Count   %Total
1       623     94.39   MOI-estimate
2       22      97.73
3       15      100.00
```

The overall estimate is computed using the percentile cutoff (--maxfact) provided by the used. The example above used the default 90% to determine a MOI of 1 (i.e. 94% of the total regions have a single MOI estimate).

4. EXAMPLE files

Download the example files from pathogenseq.lshtm.ac.uk/estmoi
extract the files using the command:
tar -vxzf estMoi.1chr.tar.gz

and run estMOI as shown in step 2 above. To re-run estMOI with a higher minimum haplotype score, see the next step.

5. RE-running estMOI

If you plan to run estMOI more than once for a single sample (eg. to optimise or test different parameters, or just to see what is in the blackbox), using the option *--tmp* is highly recommended. The temporary working directory with intermediate files will not be deleted making subsequent estimates faster. Please note that if you change parameters involving distance between SNPs, maximum insert size etc., you need to delete the temporary directory and rerunto get accurate estimates.

6. CONTACT

Please email bugs, problems and comments to samuel.assefa at lshtm.ac.uk