|  |
| --- |
| TITLE: PHE MLST software (PMS) software document |

|  |
| --- |
| **INTRODUCTION**  PMS is modified version of SRST (Short read sequence typing) software.[1] PMS uses short read Illumina fastq files to call MLST profiles. The software uses mapping and SNP calling tools to assign MLST profiles to bacterial genomic sequence data.  The software maps each readset to all possible locus variant sequences and determines the most likely allele at each locus.[1] An allele is assigned if the readset matches 100% to the locus variant sequence with zero SNP/INDELs. Where an exact match to a known allele cannot be found (SNP/ INDELs detected in all alleles), the closest allele is reported. The software then calculates a coverage statistic scores for each designated allele to assess the quality of the match. Low coverage is the most common reason for failure of allele assignment and this should be evident from coverage statistic score values.[1]  PMS software generates two output files, describing allele designation and associated coverage metrics for each assignment, namely <sample ID>.results.xml and <sample ID>.MLST\_result.csv files.  Detailed information about the <sample ID>.results.xml and <sample ID>.MLST\_result.csv output files are described in section 1 and 2. |

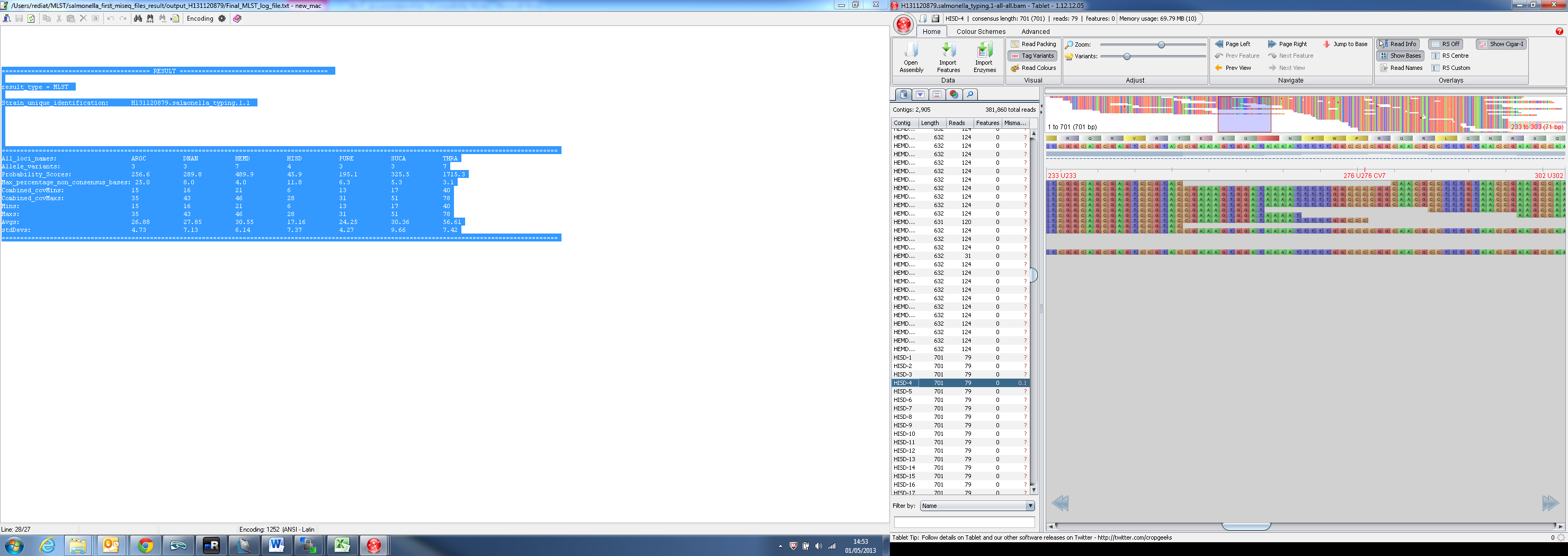
1 <sample ID>.MLST\_result.csv file

1.1 Coverage statistic metrics

Once the most likely allele is determined at each MLST locus, four coverage metrics are calculated for each MLST locus to assess the confidence of these assignments. The coverage metric values are reported in the <sample ID>.MLST\_result.csv output file.

The four coverage metrics are detailed below:

* Minimum consensus depth: the minimum number of reads that match 100% to the reference sequences at any single position. For example: the minimum number of reads that align to reference sequences (e.g. HISD-4 locus variant sequences) is 6 (See Fig 1). This value excludes non-consensus bases.



Mapped reads from tested isolate

Reference sequences (HISD-4 locus variant sequences)

*Figure 1: Display the minimum number of reads that match 100% to the reference sequences*

Min = 6

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ST value | 13 | | | | | | | | |
| Predicted Serotype | ('Agona (Achtman)', 360) | | | | | | | | |
| locus name | Allele variant | Percentage coverage | Max percentage non consensus base of each locus | Minimum consensus depth of each locus | Mean consensus depth of each locus | Number Of SNPs | SNPs Lists | Number of INDELs | INDELs Lists |
| AROC | 3 | 100 | 8 | 15 | 26.88 | 0 |  | 0 |  |
| DNAN | 3 | 100 | 4 | 16 | 27.85 | 0 |  | 0 |  |
| HEMD | 7 | 100 | 11.8 | 21 | 30.55 | 0 |  | 0 |  |
| HISD | 4 | 100 | 6.3 | 6 | 17.16 | 0 |  | 0 |  |
| PURE | 3 | 100 | 5.3 | 13 | 24.25 | 0 |  | 0 |  |
| SUCA | 3 | 100 | 3.1 | 17 | 30.36 | 0 |  | 0 |  |
| THRA | 7 | 100 | 25.0 | 40 | 56.61 | 0 |  | 0 |  |

*Table 1:* <sample ID>*.MLST\_result.csv: Details allele designation and associated confidence statistic metrics for each assignment*

* Mean consensus depth: display the average number of reads that match the reference sequences. This value excludes non-consensus bases.
* Max percentage non consensus bases: This value enables a user to identify potential mixed samples in a clinical sample.

Percentage non-consensus base is calculated for each position of the locus variant sequence by using the following formula:

Percentage non-consensus bases = Num reads mapped to reference sequence with non-consensus \* 100  
 Total number of reads aligned to reference sequence

Once the percentage non consensus bases are calculated for each position at each MLST locus, the maximum percentage non-consensus base value is determined. A max percentage non consensus bases above 15% indicates a potentially mixed sample.

* Percentage coverage: records coverage across the length of each allele. Only alleles with 100% coverage are reported.

1.2 ST values

Once the most likely allele is determined at each MLST locus, ST value is determined. The ST value is reported in the <sample ID>.MLST\_result.csv output file.

Seven different formats of ST values are returned depending on the allele designation and coverage statistic score values:

* Full profile e.g. 13

Indicates a Full MLST profile assignment with high confidence where the Minimum consensus depth is above 5.

* Full profile with low coverage e.g. \*13

Indicates a Full MLST profile assignment with low confidence where the Minimum consensus depth is equal to/below 5 or the max percentage non consensus base is equal to/above 15. The ST value is flagged by ‘\*’

* Failed(incomplete locus coverage)

Indicates incomplete locus coverage (min percentage coverage below 100%). The ST is reported as Failed(incomplete locus coverage).

* Novel allele due to presence of SNP

1. e.g NOVEL allele. Closest ST:5 (SLV)
2. e.g NOVEL allele. Closest ST:10 (DLV)

Indicates that one or more loci did not match any existing alleles. In this case, the closest ST type is indicated and the number of locus variants from this ST included within parenthesises. (e.g. NOVEL allele. Closest ST:5 (SLV) indicates a precise ST could not be detected, but the closest match is to ST5, from which the data set differs by one locus

1. e.g. NOVEL allele. cannot determine closest ST (SLV)

Indicates that one or more loci did not match any existing alleles but that in this case, a closest ST type cannot be determined.

* Novel ST profile

Indicates novel combination of alleles is detected. The ST is reported as:

1. Novel ST. Closest ST: closest\_match: 19 (SLV): Indicates a novel ST that is a single locus variant of known STs (19)

2. NOVEL ST. (no SLV)

1.3 Predicted Serotype (ONLY for salmonella samples)

Once the ST value is determined, predicted serotype and the number of occurrences associated with the ST in the PHE/Achtmann database is reported in the <sample ID>.MLST\_result.csv output file.

1. Serotype numeral: The predicted serotype and the number of occurrences associated with the ST in the PHE/Achtmann database. e.g.: (Enteritidis, 208), (Rosenberg, 2), (Nitra, 2), (Moscow, 1)

2. no ST-serotype numeral: No serotype is predicted for this ST in the PHE/Achtmann database. e.g (no ST-serotype, 1)

3. no ST-serotype: Due to a NOVEL\_ST or NOVEL\_allele it is not possible to query the PHE/Achtmann database for a predicted serotype.

1.4 Novel allele

If a locus does not match any existing locus variants, the closest allele is returned and flagged by ‘\*’ (e.g.\*55) and the SNP position between the allele in the tested isolate and the closest allele is reported in the <sample ID>.MLST\_result.csv output file.

e.g. locus:\*55 , SNP-position:90, reference base: C, number of mapped reads: A(1) C(2) G(0) T(163). This indicates a SNP at position 90, with 163 reads reporting a ‘T’ at this position See Figure 2. Note: 100 base flanking sequences are concatenated to the reference sequences.



SNP position

Reference sequences, spi-55.fa

Substitution at position 90 (190-100)

*Figure 2: The difference between the tested isolate and closest allele sequence. e.g the different between spi allele in the tested isolate and spi allele 55 is a substitution at position 90 (C-T).*

2. <sample ID>.results.xml output file

The <sample ID>.results.xml output file also reports coverage statistic metrics and ST value. In addition, it reports ‘Traffic light system’ and software version number.

* NGS sample ID : sample identifier
* Workflow value : tested isolate, e.g. salmonella-typing
* Version : software version number
* MLST value : ST value
* Profile: allele variant for each locus
* QC mean consensus depth: Single value representing the minimum average consensus depth of all loci
* QC mean consensus depth of each locus: An average of the consensus depth across the full length of each locus. List of multiple values, one for each locus.
* QC max percentage non consensus base: The largest of the maximum percentage non consensus base values from allloci
* QC max percentage non consensus base of each locus: The maximum percentage non consensus base valuesacross the full length of each locus. List of multiple values, one for each locus.
* QC minimum consensus depth: Single value representing the minimum consensus depth of all loci
* QC minimum consensus depth of each locus: Minimum consensus depth across the full length of each locus. List ofmultiple values, one for each locus.
* QC percentage coverage: Percentage coverage across allele length. NB: This value should always be 100%
* Predicted\_serotype: Predicted serotype and the number of occurrences associated with the ST in the PHE/Achtmann database. (for ONLY salmonella sample)
* Traffic light system: The “MLST typing software” validates the results based on coverage metrics and writes a cut-off value standard based on the “Traffic light system. The “Traffic light system” is assigned based on the following cut-off values:

1. The “GREEN traffic light” indication is assigned if the :

“max percentage non consensus depth” < 15% and

Complete pileup= “TRUE” and

“Minimum consensus depth” > 2 and

“Percentage coverage” =100 and

ST not "Failed(incomplete locus coverage)

1. The “RED traffic light” indication is assigned if the :

Complete pileup= “FAIL” or

“Percentage coverage” < 100 or

` ST is "Failed(incomplete locus coverage)

1. The “AMBER traffic light” indication is assigned if the there is no exact fit which matches either GREEN or RED

Example of results.xml output file

Example of results.xml output file

<ngs\_sample id=" <sample ID>">

<workflow value="salmonella-typing" version="1-0"/>

<results>

<result type="MLST" value="13">

<result\_data type="profile" value="3,3,7,4,3,3,7"/>

<result\_data type="QC\_minimum\_consensus\_depth" value="6"/>

<result\_data type="QC\_minimum\_consensus\_depth\_of\_each\_locus " value="15,16,21,6,13,17,40"/>

<result\_data type="QC\_max\_percentage\_non\_consensus\_base" value="25"/>

<result\_data type="QC\_max\_percentage\_non\_consensus\_base of\_each\_locus " value="8,4,11.8,6.3,5.3,3.1,25.0"/>

<result\_data type="QC\_mean\_consensus\_depth" value="17.16"/>

<result\_data type="QC\_mean\_consensus\_depth\_ of\_each\_locus " value="26.88,27.85,30.55,17.16,24.25,30.36,56.61"/>

<result\_data type="QC\_percentage\_coverage" value="100"/>

<result\_data type="QC\_complete\_pileup" value="TRUE"/>

<result\_data type="QC\_traffic\_light" value="GREEN"/>

<result\_data type="predicted\_serotype" value=" Agona', 15)"/>

</result>

</results>

</ngs\_sample>

Reference

# 1. [Inouye M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Inouye%20M%5BAuthor%5D&cauthor=true&cauthor_uid=22827703), [Conway TC](http://www.ncbi.nlm.nih.gov/pubmed/?term=Conway%20TC%5BAuthor%5D&cauthor=true&cauthor_uid=22827703), [Zobel J](http://www.ncbi.nlm.nih.gov/pubmed/?term=Zobel%20J%5BAuthor%5D&cauthor=true&cauthor_uid=22827703), [Holt KE](http://www.ncbi.nlm.nih.gov/pubmed/?term=Holt%20KE%5BAuthor%5D&cauthor=true&cauthor_uid=22827703). (2012) Short read sequence typing (SRST): multi-locus sequence types from short reads. [BMC Genomics.](http://www.ncbi.nlm.nih.gov/pubmed/22827703) 13:338.