Validation of MLST assignments for PATH-SAFE

Purpose

This report is intended to demonstrate the accuracy of the MLST assignment implementation at Pathogenwatch for use with the PATH-SAFE programme. In agreement with the technical community advisory group, the Achman 7 gene MLST scheme hosted by EnteroBase was chosen as the standard for PATH-SAFE. The primary aim of this validation is to show that Pathogenwatch will assign the correct MLST code when using the PATH-SAFE assemblies; the secondary aim is to demonstrate the Pathogenwatch software produces identical assignments to EnteroBase given the same assembly.

MLST profile construction

An MLST profile consists of a defined number of non-overlapping loci, usually seven. For each locus, a single code representing a specific allele is reported (an “ST” code), while the combination of alleles is also reported using a unique identifier. For each MLST scheme, a single provider maintains the allele and combined ST codes, and assigns a positive integer code for each unique entity. More recently identified alleles will have larger integer codes.

When searching a genome against a scheme, it is possible for there to be cross-hits or overlaps between potential locus matches, as well as duplications either from assembly error or genomic rearrangement. Firstly only a limited overlap is allowed between matches, typically less than 30 nucleotides. If there are multiple matches to different regions (e.g. paralogs) then exact matches to known alleles are selected in preference for the MLST profile. If there are multiple known alleles present, then the one with the oldest (lowest integer) ST code is selected as the representative. If there is no exact match then the matches are sorted according to match length and then sequence identity against the known alleles and the top hit chosen as the (novel) allele. If there are still two equally plausible matches the one close to the oldest allele is selected.

Process

The Pathogenwatch MLST tool was updated to use the EnteroBase MLST scheme dated 23rd November 2023. Assemblies generated using EToKi, as according to the assembly pipeline validation, were searched using the Pathogenwatch tool to produce the MLST assignments. The corresponding record was identified in EnteroBase and the expected MLST assignments extracted. If the assigned ST code did not agree with the expected ST code then the original FASTA at EnteroBase was downloaded and searched directly. The exception is the EURL dataset for which the expected MLST assignment was provided by the ECDC.

Tested versions

MLST scheme	23/11/2023
Pathogenwatch MLST	v5.3.0

Tested resources

Name	**# genomes**	**Notes**
EURL references	173	EU standard for Salmonella serovars
Assembler validation	188	The set of genomes produced by the assembler pipeline validation process by EnteroBase.
Reference genomes	983	Complete genome sequences.
SSSCDRL	513	A broad surveillance data set from the SSSCDRL.

Results

Overall summary

Name	**Total**	**ST agrees**	**Differences**	**Notes**
APHA serotyping	30	30	0	
EURL	173	172	1	Likely to be assembly error
Reference	787	784	3	Difference in assemblies. In 2 cases, an ST is assigned.
Assembler validation	188	186	2	In both cases an ST is assigned
SSSCDRL	513	503	9	7 were missing a locus, while 2 had a novel allele for a single locus.

Reference genomes

The initial comparison identified three differences between the STs produced by Pathogenwatch and the STs assigned by EnteroBase across the 787 genomes for which EnteroBase records could be found. Ultimately, two of these differences were due to the tested assembly being an update on a previous one, and there is a substantial difference in the genome (GCA_000487295 & GCF_000272895).

EnteroBase will assemble the first set of reads for a sample, but then will not necessarily update the assembly if it is resequenced, nor update the corresponding MLST or serotypes. In this case the change in genome is so dramatic, including a very different serotype, it is likely that a sample has been swapped. In both cases the updated genome is assigned an ST and a full set of allele codes, implying the updated genomes are valid. Also, searching

the EnteroBase assemblies with the Pathogenwatch MLST tool produces the same original annotations as EnteroBase.

The third conflict (GCA_002234475) is a result of how novel alleles are handled by the two systems. The *sucA* allele for this genome has not yet been assigned a code by EnteroBase. When there is a novel allele, Pathogenwatch will assume there is no ST assigned, but in this case EnteroBase appears to have provided a permanent ST code prior to the assignment of the allele code. So, in essence, both systems are providing a valid answer here.

Assembler validation

Across the 188 genomes, initially 2 differences were identified (SRR1544221 & SRR22891344). Both cases were resolved as for the two reference genome conflicts, and were due to samples with updated reads that produced different MLST assignments. Again the new assignments had full ST codes and allele codes, and searching EnteroBase assemblies produces the same MLST assignment as EnteroBase.

EURL validation

A single error was identified here for genome 19I. The *aroC* gene should have been assigned allele 454, but instead was identified as a novel allele and novel code assigned, leading to a difference in the ST. This difference is due to a variation at that locus in the assembly; however without detailed examination of the raw read data it is not possible to determine if the mutation is real and therefore providing a novel allele assignment is correct, or whether the assembler has made an error and introduced a sequence variant.

APHA serotyping

All 30 assignments were identical.

SSCDRL

There were 9 differences found out of 513 genomes. In seven cases a locus was missing, while in two cases an novel allele ST was called instead of the assigned EnteroBase code. In all nine cases, searching against the EnteroBase versions of the assemblies produce the same ST code as on EnteroBase.

Conclusion

In all cases, there was a high concordance between the expected and assigned ST codes, suggesting the PATH-SAFE pipeline will be able to correctly place genomes in the correct groups. Furthermore, testing against the original assemblies shows 100% concordance with EnteroBase. This gives confidence that the CGPS in-house tool is correct in general.

Appendix

Appendix A - Link to full results

https://docs.google.com/spreadsheets/d/1QpErna0FEBPwPz2E-Wbb5FWaO7acdpDXhIsyCd6k--4/edit?usp=sharing

Appendix B - Link to MLST software

https://github.com/pathogenwatch-oss/mlst