Overcoming H37Rv bias: a k-mer based approach to isolate-specific masking

D. J. Whiley[1], M. R. Domingo-Sananes[1], C. J. Meehan[1,2]

[1]Nottingham Trent University, Nottingham, UK, [2]Institute of Tropical Medicine, Antwerp, Belgium

Introduction

- Genome masking is routinely applied when mapping sequencing reads to the *Mycobacterium tuberculosis* reference genome H37Rv.
- Masking increases confidence in calling single nucleotide polymorphisms (SNPs) by preventing reads from mapping ambiguously.
- There is a need to develop new isolate-specific masking schemes to avoid single reference bias in *M. tuberculosis* data analysis.

Methods

Isolate closed (complete) genome.

↓

Align (pairwise) closed genomes to generate ground truth SNPs.

↓

Identify regions of poor mapability by mapping the isolate Illumina reads back to it (DOI:10.1093/bioinformatics/btaa222).

↓

Use these regions to create an isolate-specific masking scheme.

↓

Custom Snippy (github.com/tseemann/snippy) parameters are used to call SNPs using a non-H37Rv reference.

↓

Post-mask Snippy output SNPs.

↓

Compare masked and ground truth SNPs

Lineage-specific masking

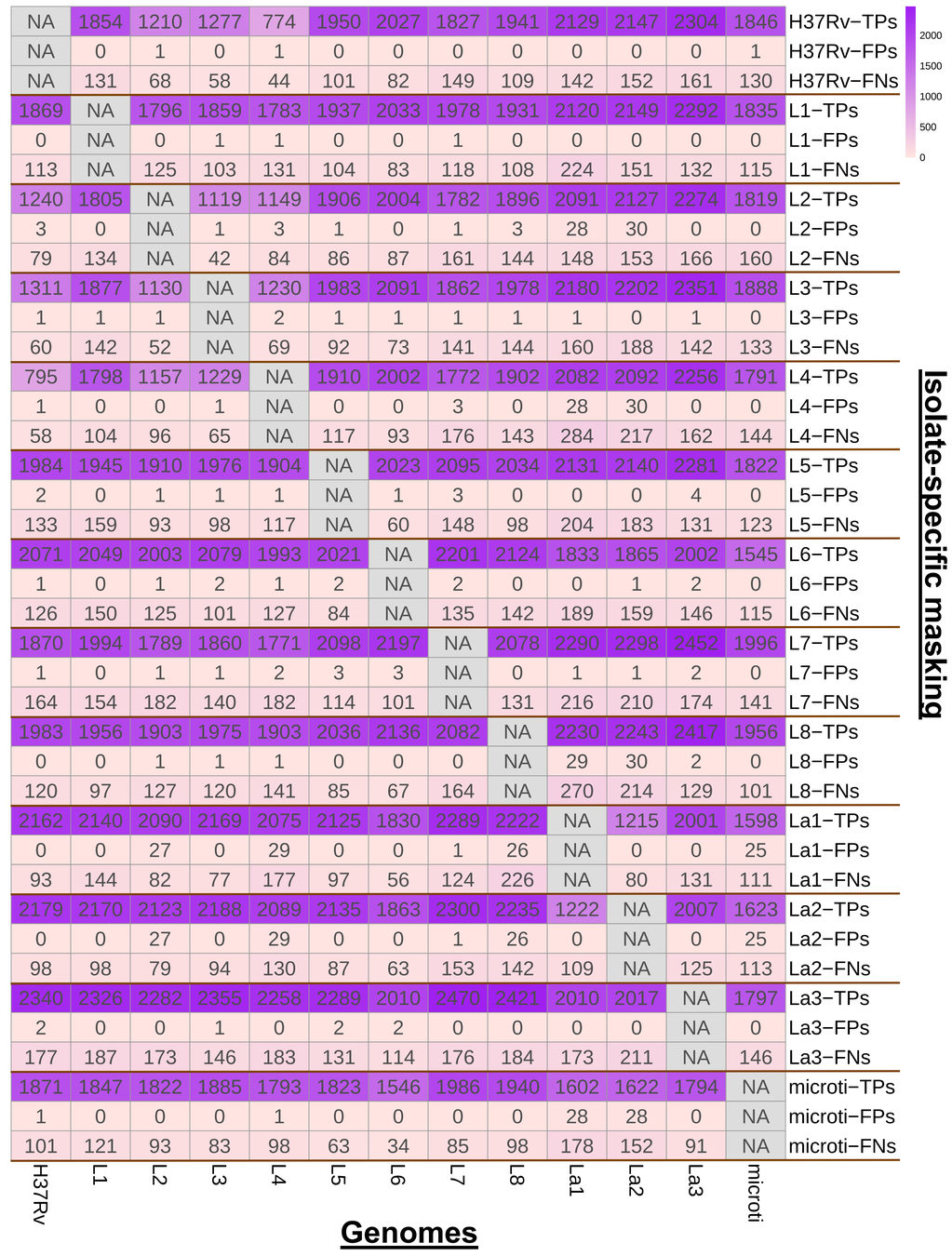

Figure 3. Summary of lineage specific masking. Heat map depicting the number of true positive (TP), false positive (FP), and false negative (FN) SNP calls using lineage-specific masking schemes. A representative (single) isolate was taken from each lineage to show how different masking schemes change the number of SNPs between other isolates.

Validation

Masking scheme:	True positive rate (Sensitivity):	True negative rate (Specificity):	False negative rate	False positive rate	Positive predictive value	Negative predictive value
This study	0.7621644	0.9999996	0.2378356	0.0000004	0.9990174	0.9998782
Marin et al (2022)	0.7360979	0.9999998	0.2639021	0.0000002	0.9993524	0.9998649
Gold standard	0.6961633	0.9999998	0.3038367	0.0000002	0.9993152	0.9998445

Figure 1. Comparison of H37Rv masking schemes. The masking scheme pipeline developed in this study was comparable to the gold standard scheme and the scheme developed by Marin et al. (2022) (DOI: 10.1093/bioinformatics/btac023). Results of all 13 isolates mapped to H37Rv for each scheme were combined and summarised.

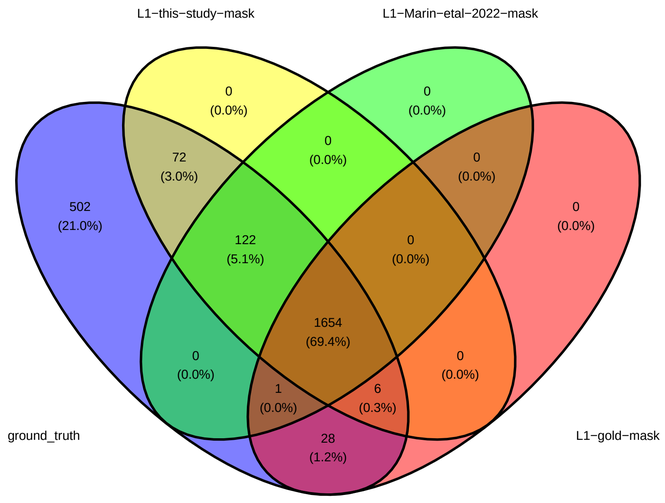

Figure 2. L1 isolate SNP recall using different masking schemes. Venn diagram depicting the number of SNPs called using different masking schemes. The masking scheme developed here was superior to the others in terms of true positive and false positive calls

Conclusions

- The masking pipeline developed in this study has a higher recall of TP SNPs compared to other masking schemes for H37Rv.
- This masking pipeline has comparable (low) false positive SNP rates compared to other masking schemes for H37Rv.
- The k-mer based mapability to genome masking can be applied to other isolates to create a robust masking scheme.
- This research paves the way for isolate-specific masking and a divergence from single reference masking.

Find Dan to discuss

Print a copy

Nottingham Trent University
School of Science and Technology
INSTITUTE OF TROPICAL MEDICINE ANTWERP