Álvaro Chiner-Oms
22-07-2022

Proposal of MTBC genomic regions to mask for phylogenetic and epidemiological studies

Classically, in our group we have always filtered out SNPs falling in PE/PPE genes, phages, repeats and the intergenic regions that flanking them (Dataset S6 on https://www.pnas.org/doi/full/10.1073/pnas.2113600119#supplementary-materials). We used to filter SNPs in these regions as many of the SNPs falling on them are homoplastic (potentially due to missmappings). These SNPs can affect phylogenetic reconstructions and inferences of epidemiological relationships between samples. We based this filter on the annotation of the SNPs called and, to my knowledge, we never evaluated in a systematic way the confidence of these regions. Applying this filter, we were able to call **89.6%** of the MTBC genome.

Recently, Marin *et al.* (https://academic.oup.com/bioinformatics/article/38/7/1781/6502279) proposed a new set of regions that are not confident for mapping. Some of these regions overlap (in different degrees) with some of the regions that we classically exclude, while some others don't. Applying this filter, they call **93%** of the MTBC genome.

So, as we are trying to update our genomic analysis pipeline for TB, we decided to formally evaluate both approaches to have an objective criteria for selecting the regions to mask for our phylogenetic and epidemiologic studies. For this analysis, we used the genomic dataset published in Coll *et al.*, 2014 (https://www.nature.com/articles/ncomms5812) which has genomes representative of the global diversity of the MTBC complex.

We applied our standard pipeline to the samples and reconstructed a phylogenetic tree. Later, we detect the number of homoplastic steps (ie. number of independent appearances in the phylogeny) for each SNP. Non-homoplastic SNPs will have 1 homoplastic step while homoplastic SNPs will have >1 homoplastic steps (Figure 1).

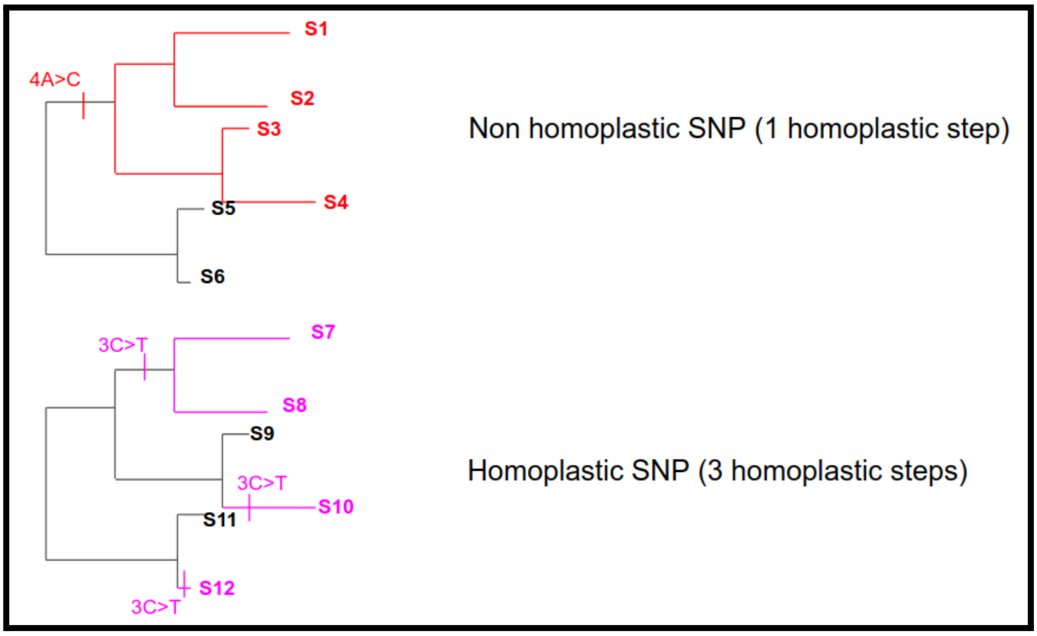

S1
4A>C
S2
S3
Non homoplastic SNP (1 homoplastic step)
S5
S4
S6
S7
3C>T
S8
S9
3C>T
Homoplastic SNP (3 homoplastic steps)
S10
S11
S12
3C>T

Figure 1. Example of some of the possible homoplastic scenarios. Non- homoplastic SNP (up) vs homoplastic SNP (down).

Next, we evaluated the homoplastic steps of the positions that passed all filters, in positions only filtered by Marin *et al.*, in the positions filtered by annotation (our initial filtering), and in the positions filtered by both approaches.

~1200 samples, MTBC wide, Coll et al. 2014

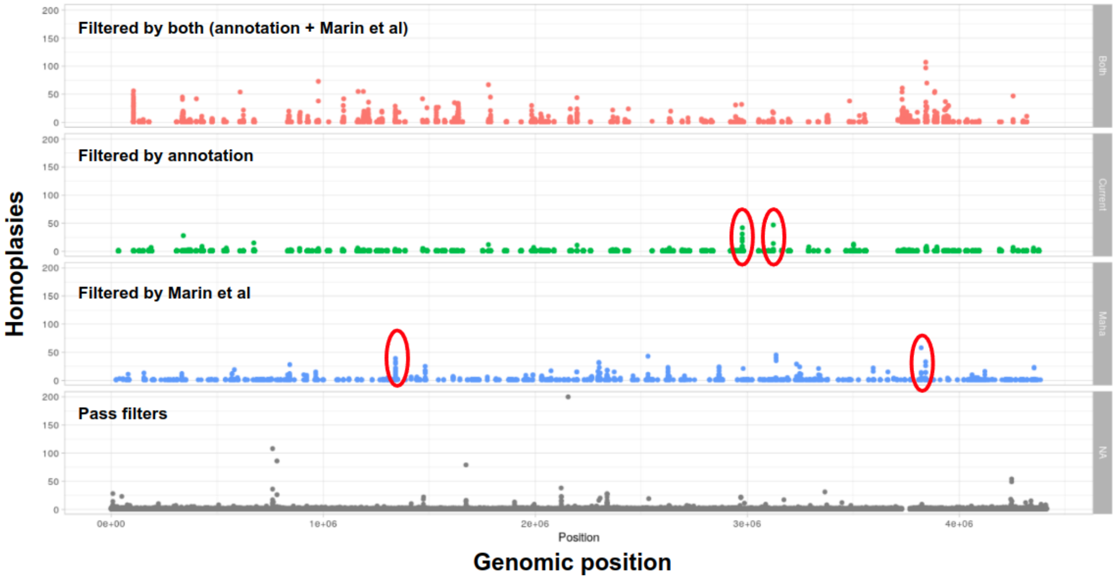


Figure 2. Number of homoplastic steps (y-axis) by genomic position (x-axis) for non filtered positions (gray), Marin *et al.* filtered (blue), annotation filtered (green) and positions filtered by both approaches (red). Red circles mark accumulation of homoplastic SNPs on positions that are either filtered by annotation or by Marin *et al.* but not by both methods.

We found that positions filtered by both methods accumulate a lot of homoplastic steps. However, there are some specific positions that are filtered by only one of both methods that also accumulate homoplastic SNPs. We acknowledge that these homoplastic SNPs may be real, and not the result of erroneous mappings. However, despite the real reason for their homoplasticity, these positions could have negative effects on phylogenetic inference of close samples and in epidemiological and clustering studies. So, we propose to filter out SNPs called in the following genomic regions:

- Those currently discarded by the annotation filter AND Marin *et al.* article.
- Those that are either discarded by annotation filter OR Marin et al methodology IF:
    - Mean number of homoplastic steps in the genomic region is > 1
    - Rate of homoplastic SNPs in genomic region is >= 0.5

With this new proposal we will cover **95%** of the genome, and skip only PE/PPE, phages, repeat and IG that Marin identifies as prone to errors, and homoplastic positions filtered by either of the methods.

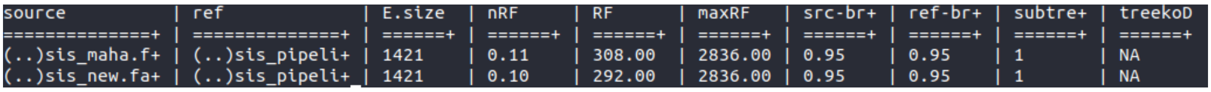
Testing the approach

If we apply those filters to the Coll *et al.* dataset we obtained the following number of SNPs:

No filter	138.555 SNPs
Annotation filter	124.307 SNPs
Marin *et al.*	130.992 SNPs
New proposal	130.914 SNPs

Our new proposal is the one that calls a higher percentage of the MTB genome (95%). However, the number of SNPs called is slightly lower than the number of SNPs called after applying the Marin *et al.* filter. This fact suggests that we are not increasing the number of false positive SNPs called, as we increase the % of genome analyzed but the number of SNPs detected did not increase.

In addition, we generated a phylogeny with the new proposed filter and with the Marin et al. filter. And we compared both phylogenies against the one derived from the annotation filter method as a reference (as this was our usual approach). We perform these comparisons using the ete3 software.



```
source          | ref             | E.size  | nRF     | RF      | maxRF   | src-br+ | ref-br+ | subtre+ | treekoD
==============+ | ==============+ | ======+ | ======+ | ======+ | ======+ | ======+ | ======+ | ======+ | ======+
(..)sis_maha.f+ | (..)sis_pipeli+ | 1421    | 0.11    | 308.00  | 2836.00 | 0.95    | 0.95    | 1       | NA
(..)sis_new.fa+ | (..)sis_pipeli+ | 1421    | 0.10    | 292.00  | 2836.00 | 0.95    | 0.95    | 1       | NA
```

Although both phylogenies match a lot the reference one (95% of edge similarity in both cases), the one obtained with the new proposal is slightly more similar to the one obtained by applying the annotation filter (second row), according to the Robinson-Foulds distance.