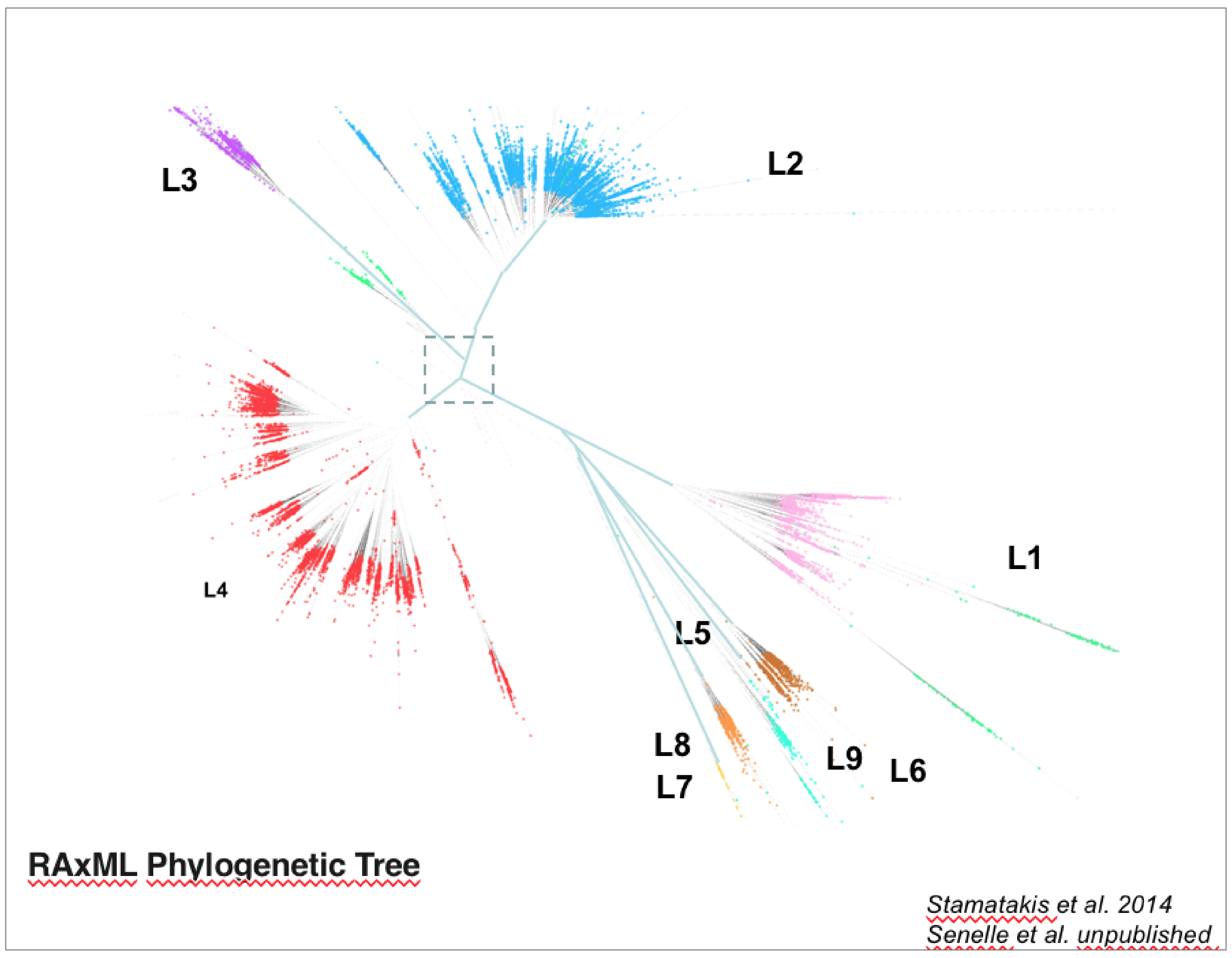
The growing chaos of tuberculosis population genomics at the era of 'Big Data': sorting out the wheat from the chaff

C. Sola[1,2], G. Senelle[3,4], MR Sahal[1,2], K. La[2,6], T. Billard-Pomares[7], J. Marin[2,8], A. Bridier-Nahmias[2], C. Guyeux[3,4], G Refrégier[1,5], E. Carbonnelle[2,7], E. Cambau[2,6]

[1]Université Paris-Saclay, 91190, Gif-sur-Yvette, France; [2]Université Paris-Cité, IAME, UMR 1137, INSERM, Paris; [3]Université Bourgogne Franche-Comté (UBFC), Besançon, France; [4]FEMTO-ST Institute, UMR 6174 CNRS-Université Bourgogne Franche-Comté (UBFC); [5]Ecologie Systématique Evolution, Université Paris-Saclay, CNRS, AgroParisTech, UMR ESE, 91405, Orsay, France; [6]AP-HP, GHU Nord site Bichat, Service de mycobactériologie spécialisée et de référence, Paris; [7]Service de microbiologie clinique, Hôpital Avicenne, 93017 Bobigny, France; Université Paris 13, IAME, Inserm, 93017 Bobigny, France; [8] Université Paris 13, IAME, UMR 1137, INSERM, Paris

Introduction

The publication of a couple of recent landmark papers (Freschi *et al.* 2021, Napier *et al*. 2021, Coscolla *et al.* 2021, Thawornwattana *et al.* 2021) claiming the discovery of new WGS-defined clades, prompted us to reevaluate both the SNP informativity and the hierarchical naming of some of the phylogenetical structures described in these articles. Thanks to a new proprietary informatical platform, ***TB-ANNOTATOR*,** we performed a benchmark analysis of these articles, and present results that allow to create new links between the pre-genomic and the post-genomic era for young researchers entering into the field, reassessing the SNP informativity, the link between polymorhic markers, and showing current discrepancies between studies, suggesting that even in large databases, the global population structure of MTBC remains strongly dependent on sample origin, WGS quality and bioinformatical tools. We also describe some recent improvements in phylogenetical analysis of MTBC.



RAxML Phylogenetic Tree

Stamatakis et al. 2014
Senelle et al. unpublished

Figure 2: RAxML-built-tree on 15901 public SRAs using the TB-ANNOTATOR informatic pipeline

In a former version with 6000 genomes, we show that L1, L5, L6, L7, L8 were sharing 3 SNPs and 288 variants with >95%, whereas L2,L3,L4,L7,L8 were sharing 0 SNPs and 0 variants, thus demonstrating the rooting of L7-L8 with L1,L5,L6,L9 branches and not with L2,L3,L4.

Bibliography:

Napier, G., *et al*. 2020. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies; Genome med, 12, 1, 114

Freschi, L *et a*l. 2021. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. Nature communications 12, 6099.

Coscolla, M., *et al*. 2021. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. Microb genomics, 7, 2, 9/2/2021

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-13

Thawornwattana, Y., *et al.* 2021. Revised nomenclature and SNP barcode for *ww*. Microbial Genomics. 7, 11.

Menardo, F. et al. 2021. Local adaptation in populations of *Mycobacterium tuberculosis* endemic to the Indian Ocean Rim [version 2; peer review: 2 approved]. F1000Research 10.

Material and Methods

15901 SRA were either downloaded from public databases (NCBI, EBI) or produced *in house*. The ***TB-ANNOTATOR*** pipeline is summarized in **Figure 1**. The Phylogenetic tree shown in **Figure 2** was produced using RaXML. ***TB-ANNOTATOR*** allows to deal not only with SNP but also with RDs, MGEs presence/absence and insertion sites.

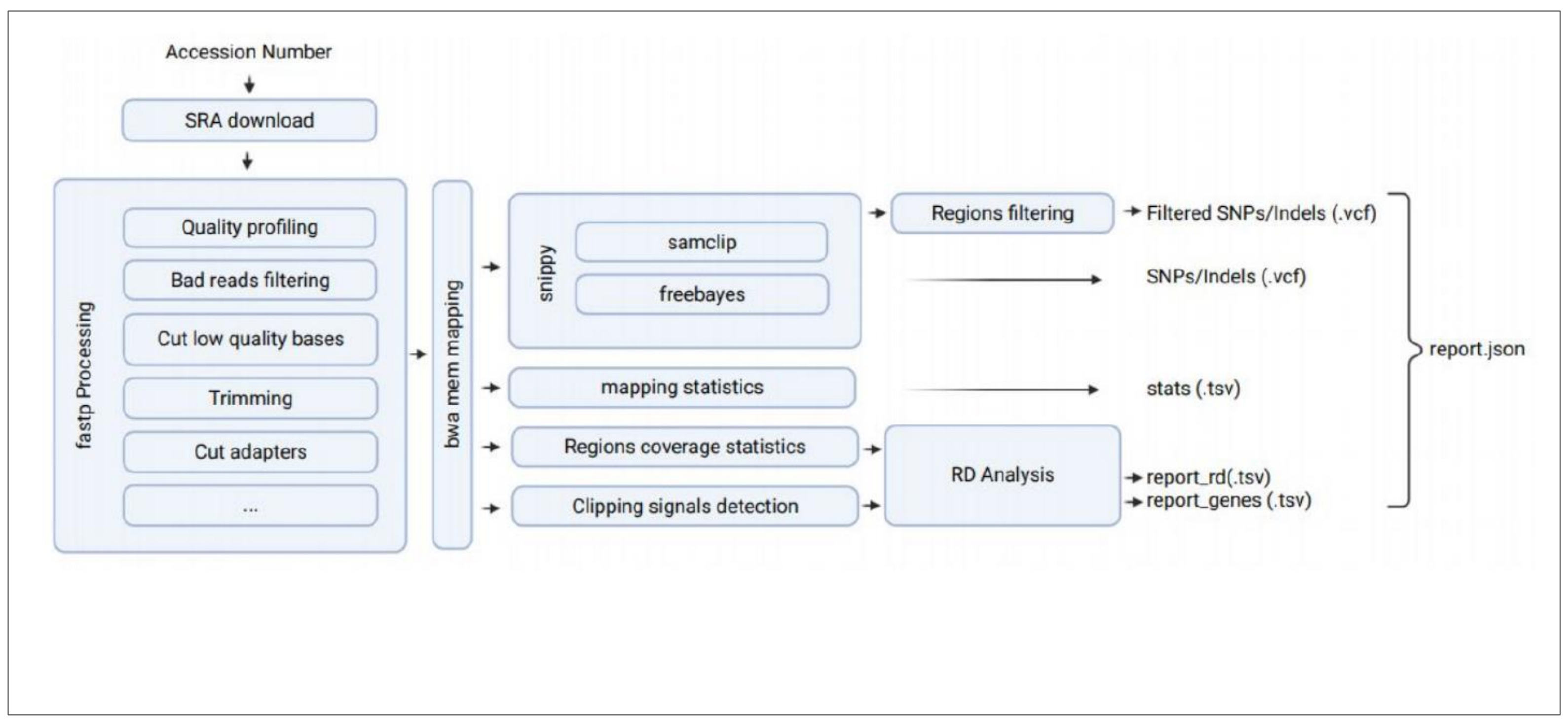


Figure 1: general algorithm of TB-ANNOTATOR

Results

The definition of meaningfull phylogenetic branches in all lineages is improved by using **TB-ANNOTATOR**. As an example, **L5 and L6** are now better defined (see also Muhammed Rabiu Sahal *et al*. poster). As another example, in **Figure 3**, **L1.1.2** is better defined by SNP position **20544** and now encompass two sublineages **L1.1.2.1** and **L1.1.2.2,** as shown below, these branches had been ignored by Freschi, Coll and Napier *et al*.

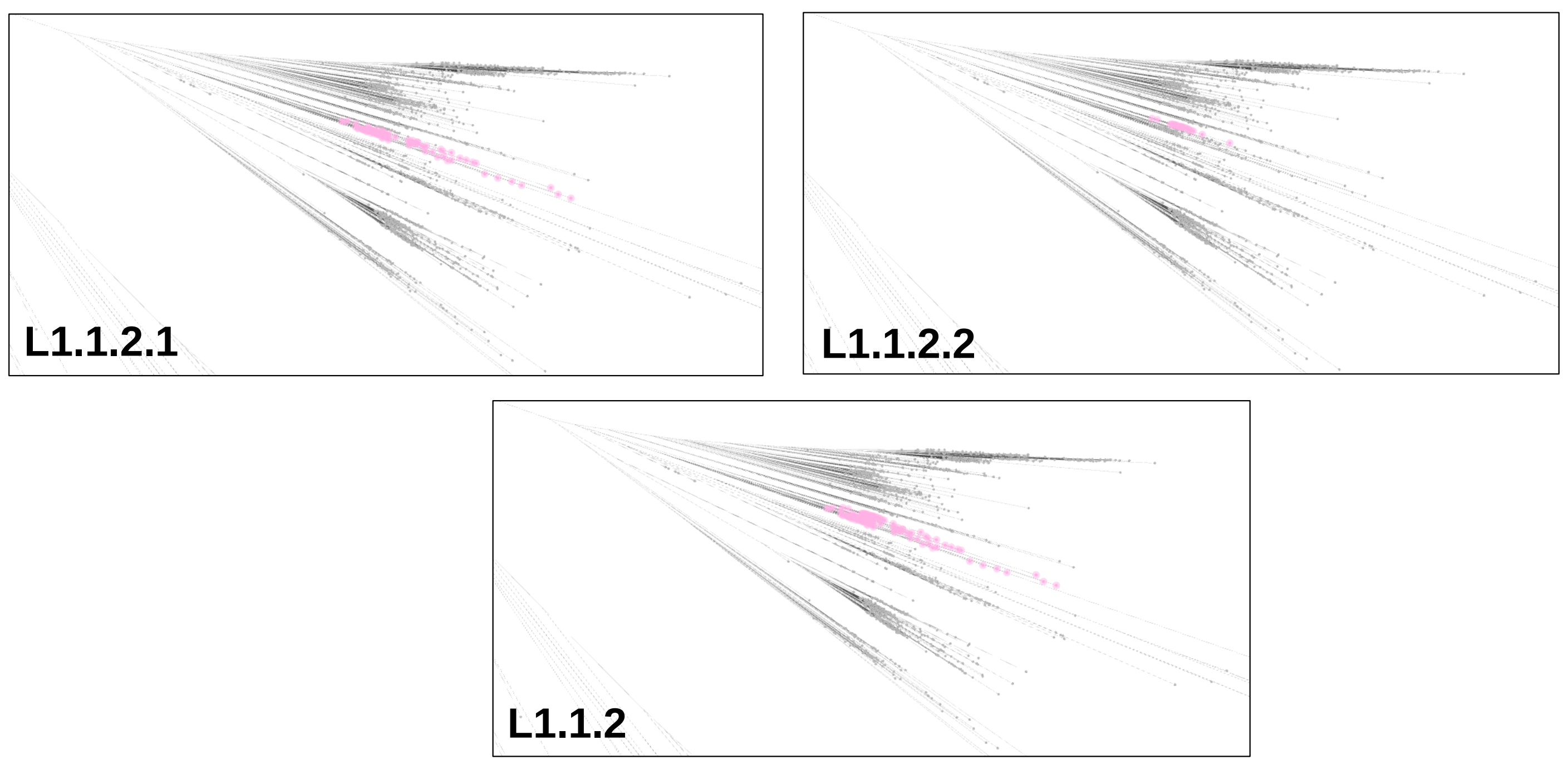


Figure 3: new definition of the L1.1.2 sublineage

In **L4, L4.5 and L4.7** structures had been ignored and **TB-ANNOTATOR allows** to gain a much deeper insight inot these families. In **L4** again, the last designated **L4.12** sublineage by Freschi was known since 2005 as « *East-Mediterranean 1* ».

Conclusion

we show by this benchmark study, that current WGS phylogenetical studies are very strongly subjected to **sampling bias** and that a stable global picture of MTBC population structure will only be achieved once a representative sample of MTBC genetic diversity will have been built. Current studies tend to preferably describe epi-linked clusters without assessing the global spatio-temporal historical picture