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

C. Sola1,2, G. Senelle3,4, MR Sahal1,2, K. La2,6, T. Billard-Pomares7, J. Marin2,8, A. Bridier-Nahmias2, C. Guyeux3,4, G Refrégier1,5, E. Carbonnelle2,7, E. Cambau2,6

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

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 RMs, MGEs presence/absence and insertion sites.

Results

The definition of meaningful 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.

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.