A proposal for Allele Validation Tool (AVT) for whole genome Multilocus Sequence Typing (wgMLST) schemas

Hugo Curado[1], Margarida Cândido[1], Mickael Silva[2], João André Carriço[2]
Francisco M. Couto[1]

[1]LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[2] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

With High Throughput Sequencing methods (HTS), sequence-based microbial identification methods have evolved to query thousands of loci present in a given bacterial species. Methods such as the ones used in BigsDB(1), are based on draft genomes (DGs) produced by *de novo* assemblies from reads, and BLAST based methodologies match the loci on DGs to existing allelic databases. However when novel alleles are found, they should be validated since the variation could be an artifact introduced by the *de novo* assembler. Thus, the role of our work is to create AVT (Allele Validation Tool), a tool that is capable of validate novel alleles found in de novo assemblers by allele calling algorithms.The proper validation of an allele can be of extreme importance, specially in the correct detection of outbreak situations based on wgMLST.

The inputs for AVT are the reads of the strains under study, the contigs produced by the *de novo* assembler and a list of all the novel alleles found. AVT uses a sequence mapper such as Bowtie2 (2), and samtools/bfctools (3), to detect any SNPs and validate coverage over the length of the allele. The output will report any SNP/indels found or if the allele is confirmed.

As future work, we plan on providing the tool as a web application and also as a standalone application (based on python and browser compatible languages) so

Hugo Curado and Margarida Cândido have worked equally on writting this article.

everyone can access it easily. Moreover, we also want to make our code open source so everyone can contribute and improve it to fit their best needs.

(1) Jolley, KA and Maiden M. *BMC Bioinformatics* 2010 **11**:595.

(2) Langmead B, Salzberg S *Nature Methods*. 2012, 9:357-359.

(3) Li H. Bioinformatics. 2011 Nov 1;27(21):2987-93.

Preference for presentation: Oral
Location: University of Minho
Author for Correspondence: hugofsvbc@alunos.fc.ul.pt

Hugo Curado and Margarida Cândido have worked equally on writting this article.