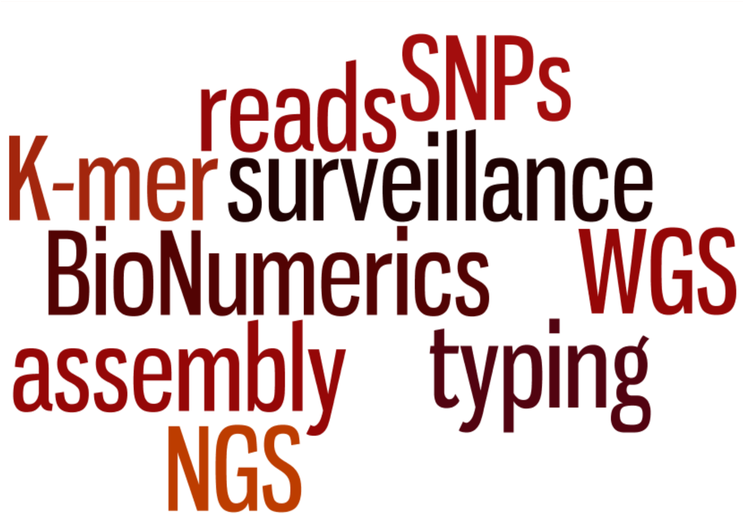
afkSNP:
assembly-free K-mer based SNP comparison of bacterial WGS samples

Jeroen Van Goey[1,*], Hannes Pouseele[1], Philip Supply[2,3] & Stefan Niemann[4]

[1]Applied Maths NV, Sint-Martens-Latem, Belgium ; [2]Genoscreen, Lille, France ; [3]INSERM, U1019, CNRS UMR 8204, Institut Pasteur de Lille, Univ Lille Nord de France, Lille, France; [4]National Reference Center for Mycobacteria, Forschungszentrum Borstel, Borstel, Germany; *info@applied-maths.com



INTRODUCTION

The democratization of whole genome sequencing (WGS) creates new possibilities for molecular surveillance of bacterial pathogens. However, easy to use bioinformatics tools to analyze these samples are often lacking:

- On a technical level: routine labs often have difficult access to the bioinformatics resources required to analyze (confidential) whole-genome sequencing data.
- On a biological level: for numerous micro-organisms there are no closely related reference sequences available, compromising the resolution of reference sequence-based approaches.

OBJECTIVES

- *Reference-free*, *assembly-free* and *fast* method to compare WGS samples based on single-nucleotide polymorphisms (SNPs).
- Tools, *integrated in the software suite BioNumerics®*, enabling fast and reliable strain comparison, with subsequent phylogenetic, statistical and data mining analyses.

METHODOLOGY

- For every base, use its flanking regions to define the location.
- For each sample, calculate the complete list of pairs of words of length k. To limit sequencing errors, we filter on
  - base quality scores,
  - overall and strand-specific coverage
  - middle base ambiguity.
- Identical repeat regions yield the same set of “locations”, middle base ambiguity filtering removes all dubious “locations”.
- For each pair of samples, the word lists are compared by counting the number of words that have the same start and end, but have a different middle base. This yields a pairwise distance matrix that can be used for further analysis.

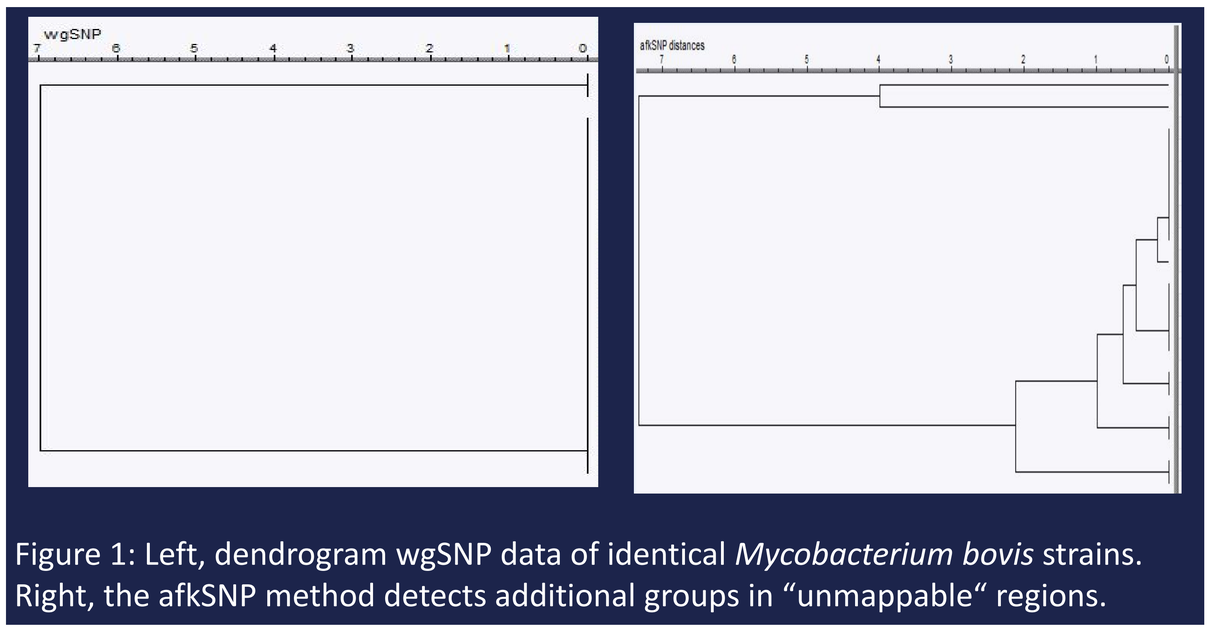


Figure 1: Left, dendrogram wgSNP data of identical *Mycobacterium bovis* strains. Right, the afkSNP method detects additional groups in “unmappable“ regions.

EXPERIMENTAL VALIDATION

- We use 22 WGS samples of identical *Mycobacterium bovis* (MBO) strains to investigate the stability and the reproducibility, and compared the results to a reference-based SNP calling procedure (wgSNP). For both methods, we found a total of 7 mutations. The afkSNP method revealed slightly more mutations in repeat regions that have been excluded from the reference-based approach due to unreliable mapping (see figure 1).
- afkSNP analysis of a set of 26 outbreak-related *Mycobacterium tuberculosis* (MTBC) samples shows the potential of the method to reveal the structure of the outbreak in a fast and objective manner. The correlation between distances calculated by afkSNP and wgSNP in this sample is 99,9% (see figure 2).

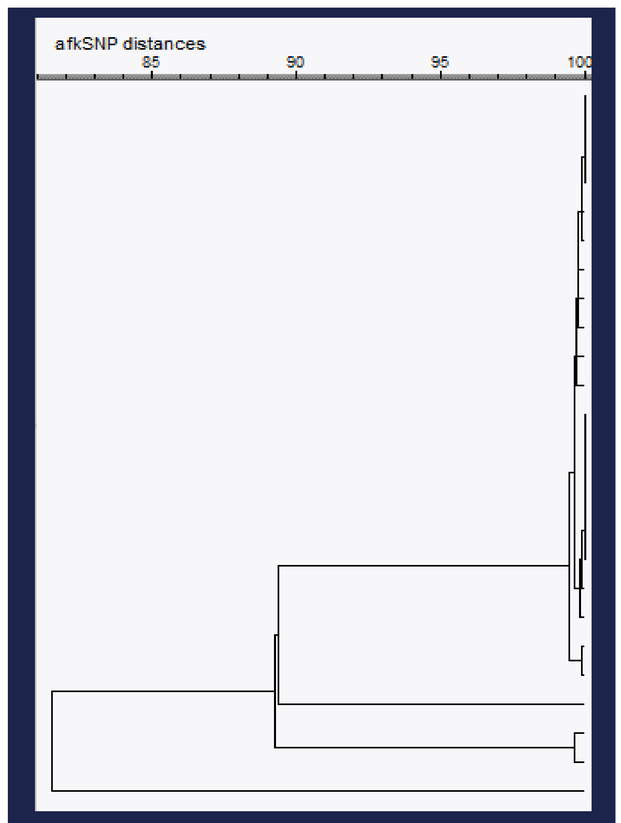


Figure 2: afkSNP dendrogram of 26 outbreak-related *Mycobacterium tuberculosis strains.* The afkSNP similarities correlate extremely well with the wgSNP data (99.9% correlation)

CONCLUSIONS

- The afkSNP method gives almost identical results as the classical SNP detection methods, but does not require a closely related reference.
- The afkSNP method compares the complete genomic content of all WGS samples, in a pairwise manner, and not just what is common with the reference sequence.
- The use of only *isolated* SNPs does restrict the analysis, but at the same time this approach avoids clusters of SNPS that have been introduced by a single evolutionary event, thus providing a less disturbed counting of evolutionary events. Therefore, the afkSNP method covers a middle ground between reference-based SNPs and whole genome multi-locus sequence typing.

ACKNOWLEDGEMENTS

All WGS data has been generated by GenoScreen in the context of the Patho-NGen-Trace project. We would like to thank Dr. Stefan Niemann for the *Mycobacterium tuberculosis* strains and Dr. Philip Supply for the *Mycobacterium bovis* strains.

Contact:
Hannes Pouseele, Applied Maths NV,
hannes_pouseele@applied-maths.com

Patho-NGen-Trace is funded by the EC under the 7th Framework Programme of the European Union