Using MALDI-TOF spectra in epidemiological surveillance for the detection of bacterial subgroups with a possible epidemic potential

Audrey Giraud-Gatineau
Aix Marseille Univ, IRD, AP-HM, SSA, VITROME, IHU Méditerranée Infection
Gaetan Texier
Centre d'Epidémiologie et de Santé Publique des Armées (CESPA)
Pierre-Edouard Fournier
Aix Marseille Univ, IRD, AP-HM, SSA, VITROME, IHU Méditerranée Infection
Didier Raoult
Aix Marseille Univ, IRD, AP-HM, MEPHI
Hervé Chaudet (herve.chaudet@gmail.com)
IHU Méditerranée Infection

Research Article

Keywords: MALDI-TOF, epidemiological surveillance, cluster analysis, epidemic

Posted Date: April 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-398130/v1

Abstract

Background For the purpose of epidemiological surveillance, the Hospital University Institute Méditerranée infection has implemented since 2013 a system named MIDaS, based on the systematic collection of routine activity materials, including MALDI-TOF spectra, and results. The objective of this paper is to present the pipeline we use for processing MALDI-TOF spectra during epidemiological surveillance in order to disclose proteinic cues that may suggest the existence of epidemic processes in complement of incidence surveillance. It is illustrated by the analysis of an alarm observed for *Streptococcus pneumoniae.*

Methods The MALDI-TOF spectra analysis process looks for the existence of clusters of spectra characterized by a double time and proteinic close proximity. This process relies on several specific methods aiming at contrasting and clustering the spectra, presenting graphically the results for an easy epidemiological interpretation, and for determining the discriminating spectra peaks with their possible identification using reference databases.

Results The use of this pipeline in the case of an alarm issued for *Streptococcus pneumoniae* has made it possible to reveal a cluster of spectra with close proteinic and temporal distances, characterized by the presence of three discriminant peaks (5228.8, 5917.8, and 8974.3 m/z) and the absence of peak 4996.9 m/z. A further investigation on UniProt KB showed that peak 5228.8 is possibly an OxaA protein and that the absent peak may be a transposase.

Conclusion This example shows this pipeline may support a quasi-real time identification and characterization of clusters that provide essential information on a potentially epidemic situation. It brings valuable information for epidemiological sensemaking and for deciding on the continuation of the epidemiological investigation, in particular the involving of additional costly resources to confirm or invalidate the alarm.

Clinical trials registration. NCT03626987

Study authorization. This study has been allowed by the French Data Protection Authority (CNIL decision DR-2018-177).

Background

Epidemiological surveillance systems have a central role in order to control and manage infectious diseases [1,2]. Since 2013, the Hospital University Institute Méditerranée infection (IHU-MI) has implemented an epidemiological surveillance system named MIDaS (for Mediterranée Infection Data Warehousing and Surveillance) made of five syndromic surveillance sub-systems. This system is based on the systematic recording of routine results issued from clinical microbiology and virology laboratories, which are not specifically done for surveillance purpose [3], including identification at species level and possibly phenotypic or genomic characters. Data from other information systems are also collected, such

as spectra files generated by the Matrix Assisted Laser Desorption Ionization – Time of Flight (MALDI-TOF) mass spectrometers used for bacterial and fungal species routine identification [4]. MIDaS automatically and systematically analyses the number of bacteria identifications in search of abnormal increases, corresponding to “surveillance alarms”. Each week during a staff an evaluation of these alarms is done in order to decide how to deal with them, doing a epidemiologic sensemaking that we have previously conceptualized under the term "situation diagnosis" [5]. During this situation diagnosis, MIDaS also helps to contextualize the alarm, allowing an “in silico” investigation based on sample and patient characteristics.

Recent publications have demonstrated that species-level surveillance alone is often insufficient to carry out the situation diagnosis [6-9] because a same bacterial species may present a great diversity of subspecies with strong variations in clinical and epidemiological expression, each of them possibly being an epidemics [10]. The search for specific genetic markers or the use of antibiograms make it possible to detect this kind of subspecies outbreaks, but requires sometime extensive extra works.

Nowadays, MALDI-TOF MS is used in routine bacterial identification and for retrospective epidemiological investigations [10-14]. It appears to be an answer to these more time-consuming and tedious laboratory techniques. Retrospective studies based on spectra clustering revealed the proteinic similarity of strains sharing the same geographical area and the same epidemic features [11-14] or the dissimilarity between epidemic strain and usual species spectra[10]. However, to our knowledge, no study reports the use of MALDI-TOF MS during a routine epidemiological surveillance activity [5,15].

The objective of this paper is to present a pipeline for processing MALDI-TOF spectra during epidemiological surveillance in order to disclose a latent clustering of a species, which may suggest the existence of epidemic processes. This description will be illustrated by the analysis of an alarm observed for *Streptococcus pneumoniae*.

Materials And Methods: Description Of The Pipeline

The spectra-based surveillance system relies upon a microbiological surveillance system associated with a MALDI-TOF MS database. The overall process flow is described in Figure 1.

Microbiological surveillance system: BALYSES subsystem

The IHU-MI is integrated in the Assistance Publique – Hôpitaux de Marseille (AP-HM) in Marseille, France, and is the unique bacterial clinical microbiology laboratory of its 4 public and university hospitals. The laboratory activity has been monitored since February 2014 by an automated surveillance system named BALYSES [7] (Bacterial real-time Laboratory-based Surveillance System), which is one the five MIDaS subsystems. Connected to the laboratory information system, this surveillance system is based on a dedicated data warehouse gathering microbiological analysis results (sample id, requesting department, date, sampling, analysis, result, possible antibiotic susceptibility testing, possible antibiotic resistance phenotype, bacterial co-identifications) and patient-related information (anonymized patient id, age, sex,

home postal code, anonymized hospital stay id, department stay date, death). It allows a systematic weekly detection of outbreaks for all bacterial species included in the database using CUSUM algorithms [16], the monitoring of trends for the sampling activity of the 15 most frequent bacterial species, and tracking of rare or new bacterial species.

MALDI-TOF MS database from IHU-MI

Since 2014, February 1st, the MALDI-TOF database has gathered around 900,000 spectra performed at the IHU-MI for the routine bacterial identifications of all patients hospitalized in the AP-HM.

The routine bacterial identification of the laboratory relies on strain culturing on blood or chocolate agar depending on the species and stopped in the middle of log phase (BioMérieux's Columbia with 5% sheep blood agar, and Becton, Dickinson and Co's Chocolate Agar GC II Agar with IsoVitaleX™ enrichment). From the cultures, a single colony is directly applied in on 2 or 4 spots on ground steel targets, air dried, overlaid with α-cyano-4-hydroxycinnamic acid matrix solution in 50 % of acetonitrile and 2.5 % of trifluoroacetic acid and air dried following the agreed protocol. All bacterial spectra are acquired using 3 Bruker Daltonics Microflex MALDI-TOF MS with FlexControl Software, using the default settings (positive linear mode within the m/z range of 2 to 20 kDa, laser frequency 60 Hz; ion source 1 voltage, 20 kV; ion source 2 voltage, 16.7 kV; lens voltage, 7.0 kV), and 240 laser shots at 60 Hz. Culture standardization is required for allowing spectra comparability within a same species. Bruker BioTyper® software allows the comparison between the spectrum and a reference database and leads to the bacterial species routine identification when the score threshold is ≥ 2.0. The Bacterial Test Standard (BTS) which is a solution of *Escherichia coli* DH5 alpha with two additional proteins, is used as a positive control and the matrix solution as a negative control for identification. Automata calibration is regularly performed as described by Bruker's protocol using the BTS.

All MALDI-TOF MS spectra ('fid' files) and their parameter files ('acqu' files) produced during the identification process are extracted from automata and saved in a specific file system storage, associated with the surveillance data warehouse.

MALDI-TOF MS analysis

Our spectra processing platform is based on a homemade program written in R [17] and mainly using the following packages: MALDIquant v1.16.2 [18] for spectra reading and quantitative analysis, seriation v1.2-2 [19] for dendrogram ordering, and BinDA v1.0.3 [20] for the protein peak discriminant analysis using binary predictors.

Spectra selection

For investigating an alarm, the related surveillance database records and their associated spectra are selected using the appropriate request (e.g. based on species, dates, antibiotic susceptibility, home location, hospital department…), with a time window extension (over a maximal period of 4 months) for

including a sufficient pre-alarm contrasting material. This delay may be shorter if the number of spectra is too huge to make the clustering readable, as for *Escherichia coli* or *Staphylococcus aureus*. The limit usually used in this case is about 1,500 spectra. During the selection process spectra quality is taken in account: only spectra of sufficient quality in terms of saturation and noise [21] and with plate controls (BTS) required for spectra deviation correction (as described below) are included in the analysis.

Spectra processing

The selected spectra are imported into the analysis platform and are then injected into a 4-step workflow, which is described below. The spectra processing includes normalization [18], double alignment of spectra [22], Main Spectrum Profiles (MSP) and intensity matrix building. During these steps, the signal to noise ratio (SNR) was 2 and was used as a peak detection threshold, the peaks with a SNR<2 being considered as noise.

As described by Gibb and Strimmer [18], the normalization is made of intensity transformation (square root method), smoothing (moving average with half window size 12), baseline correction (Statistics-sensitive Non-linear Iterative Peak-clipping algorithm, 100 iterations) and intensity recalibration (on the maximal intensity peak).

The 8 reference peaks (3637.8, 5096.8, 5381.4, 6255.4, 7274.5, 10300.1, 13683.2, 16952.3 Da) of the BTS, required for each target plate, are used for a first alignment (quadratic warping function) aiming at controlling automata-dependant drift. Spectra with reference peaks out of the built-in Microflex tolerance window (300ppm) are dropped. Using the species typical peak composition described in our panspectrome database [22], a second alignment of the spectra based on their species-specific common peaks is then done (quadratic warping function with 0.005 tolerance).

Technical replicates are averaged into main spectrum profiles (MSP), and species specific common peaks are removed in order to increase the contrast between these spectra, which belong to the same bacterial species [22].

An intensity matrix, describing the intensity of spectra peaks for each MSP, and built as recommended by S. Gibb [20], is the final deliverable of this process.

Spectra clustering

The next step is the hierarchical clustering of the intensity matrix using Bray-Curtis distance and Ward agglomeration with ordination (or seriation). The ordination is based on the Gruvaeus-Wainer method [23], which orders the leaves at each merging step such the leaves at the edges of each cluster are beside the more similar ones, ensuring the unicity of the dendrogram. Time distances between dendrogram leaves are also calculated during this step.

The results of this clustering step are presented using 2 specific graphics: a time-heated dendrogram and a time-protein double proximity heatmap. Their aim is to support epidemiological inference based on the

MSP closeness in terms of proteinic and temporal distances, suggesting the possible epidemiological relations between isolates, as elaborated by Sintchenko et al. [24]. In the time-heated dendrogram, each leaf label is coloured with a heat scale according to the case occurrence time. More the case is recent and more the color is “hot”, from blue to red. Isolates possibly belonging to a same epidemiological event are represented in the dendrogram by subtrees with labels showing the same colour. The time-protein double proximity heatmap combines a first half-matrix showing proteinic distances with a second half-matrix coloured in accordance with the time distance between MSP (Figure 4, cluster A). The double heatmap is a possible alternative illustration where groups of MSP with close proteinic-temporal distances appear as hot colour squares along the matrix diagonal.

Spectra characterization

Characterization of MSP belonging to a group is done by contrasting this group against the other MSP with a discriminant analysis on protein peaks. For this purpose, we rely on the Gibb and Strimmer’s method for differential protein expression and prediction based on binary discriminant analysis (BinDA) [20]. This method dichotomizes the intensity vector of each peak using the maximisation of the Kullback-Leibler divergence, before finally ranking them according to their discriminating power. All top-ranked peaks are automatically checked against the UniProt database (http://www.uniprot.org/) using its representational state transfer (REST) programmatic access. A mass fluctuation of ± 2 Da is allowed for the matching. For each top-ranked peak, prediction errors for group separation are estimated using cross-validation procedures [25].

Ethic information

This study has been allowed by the French Data Protection Authority (CNIL decision DR-2018-177), and declared on ClinicalTrials.gov Protocol Registration and Result System (id: NCT03626987).

Results

Surveillance system activity

At the date of February 2020 (316 weeks since 2014, February 1st), the microbiological surveillance database includes 287,679 bacterial identifications for 559 different species, from 237,196 clinical samples and 100,729 patients (137,625 hospital stays). The associated MALDI-TOF datawarehouse gathers 929,740 MALDI-TOF MS spectra. The database increases at a weekly rate of about 12,000-13,000 samples and 1,000 bacterial identifications for 1,000-1,300 new patients. The three most represented sample are urine samples (79,528 samples, 33.5% of the total) followed by blood samples (43,188 samples, 18.2%) and respiratory samples (25,966 samples, 10.9%). The ten most identified bacterial species are *Escherichia coli* (61,734 strains, 21.5% of the total), *Staphylococcus aureus* (46,791 strains, 16.3%), *Staphylococcus epidermidis* (22,180 strains, 7.7%), *Pseudomonas aeruginosa* (19,789 strains, 6.9%), *Klebsiella pneumoniae* (18,722 strains, 6.5%), *Enterococcus faecalis* (12,723 strains, 4.4%), *Enterobacter cloacae* (10,091 strains, 3.5%), *Streptococcus agalactiae* (8,124 strains, 2.8%), *Gardnerella*

vaginalis (7,773 strains, 2.7%) and *Proteus mirabilis* (5,934 strains, 2.1%). *Streptococcus pneumoniae* is involved in 2,509 strains (0.9%).

Illustrative alarm analysis

BALYSES surveillance system found an abnormal increase of *Streptococcus pneumoniae* identifications from January 31th to February 9th 2020 (5-6th weeks), with 17 cases for 12 expected. *S. pneumoniae* is known to be amongst the worldwide leading cause of death due to infectious diseases [26]. For the purpose of the MALDI-TOF investigation of this alarm, and following the protocol described above, we have considered all qualified spectra from October 1st, 2019 to February 16th, 2020 (Figure 2). The 17 patients having caused the *S. pneumoniae* alarm were 13 men and 4 women. Their mean age was 35.6 years, and the length of their hospital stay was 6.2 days in average. *S. pneumoniae* was mostly identified in blood cultures (N = 7, 41.2%), respiratory samples (N = 6, 35.3%) and deep samples (N = 2, 11.8%).

During the extended period considered for this analysis, 213 *S. pneumoniae* identifications were performed, corresponding to 171 samples from 136 patients and 138 hospital stays. A total of 644 MALDI-TOF spectra were associated in the spectra datawarehouse. After application of quality criteria, 421 (65.4%) spectra related to 123 patients (125 hospital stays) were retained for further analysis. During the spectra processing, these spectra were grouped in 125 MSPs (Main Spectrum Profile), producing an intensity matrix of 125 rows (MSPs) and 152 columns (peaks) as final result.

The results of the spectra clustering phase are presented in Figure 3 and 4. Due to the color code used for the representations, patients involved in the surveillance alarm are included in dendrogram's red labels. They are mainly concentrated in 2 subtrees (subtrees A and B), which may be associated to two simultaneous epidemiological events.

The subtree A gathers 7 MSPs produced during the previous 3 weeks, while the subtree B is a grouping of 10 MSPs produced over a 4 months period, and less pertinent for the alarm investigation. The related surveillance database data show that all MSPs of subtree A are coming from different patients, with 4 associated to patients involved in the alarm, and 4 MSPs coming from respiratory samples, 2 from deeper samples and 1 from skin sample. The time-protein double proximity heatmap (Figure 4) confirms the epidemiological interest of subtree A, showing a corresponding 'hot' square.

We have tried to find what peaks were able to contrast the MSPs of subtree A with the rest of the dendrogram, using a binary discriminant analysis (Figure 5). In this representation, a positive t-score indicates the presence of the peak and a negative t-score its absence. The best top-ranked peaks are in the 5-8 kDa bandwidth, and the 4 top-ranked are the most discriminant. Subtree A is indeed characterized by the presence of three of them (5228.8, 5917.8, and 8974.3 m/z) and the absence of peak 4996.9 m/z. Automatic checking of these 4 peaks against UniProtKB retrieved all of them (supplementary material), showing that peak 5228.8 is possibly an OxaA protein and that the absent peak may be a transposase.

Discussion

The objective was to present a pipeline using MALDI-TOF spectra in the early stages of the situation diagnosis in order to disclose a temporo-proteic cluster that could suggest the existence of an epidemic chain as suggested by Sintchenko [24]. A previous study on *Staphylococcus saprophyticus* [14], allowed us to explore the capability of MALDI-TOF MS spectral clustering in epidemiology with the identification of a particular subspecies circulating in Marseille. From this attempt, we have progressively improved the stability and power of spectra analyses with a better control of the intra and inter automaton variations (additional alignment on the BTS peaks), less analysis noise (exclusion of core peaks), the adding of visualization cues by graphical representations contrasting homogeneous temporo-proteic clusters, and the identification and characterization of discriminant peaks. All these processes are possible because this pipeline is directly connected to a single system MIDaS that systematically collects and concentrates all the data from the microbiology laboratory, both the biological results associated with patient and sample information and the MALDI-TOF spectra.

The carriage of a bacterial species in a human population is made of the cohabitation of a multitude of lineages corresponding to multiple chains of transmission. Each of them may have its own epidemiological characteristics [27-29]. This explains why genetic fingerprinting techniques such as whole-genome sequencing (WGS) are increasingly used in many epidemiological contexts, in particular for confirming that samples belong to a same epidemic chain[30] or for studying the dynamics of epidemics [8,31-33]. We cannot ignore the fact that a same genome may have different phenotypic expressions, and conversely [34]. However, in the context of our study, we hypothesize that the phenotypic expression of a strain is a proxy for its genetic profile, insofar as its culture conditions are standardized (i.e. the environmental pressure being the same during bacterial growth). By extension, we consider that a set of bacterial strains presenting a same phenotypic expression can be sufficiently similar for belonging to a sample of the same epidemic process, and, with respect to the limitations presented above, a possible epidemic clone. The aim of this pipeline is not to do subspecies identification, but to find a possible suspicion of bacterial subspecies spread and so to pre-investigate a possible outbreak by inexpensive and routinely useable methods. To fully confirm these results, genetic or molecular methods remain needed and would be done during the epidemiological investigation.

The use of this new-generation pipeline in the case of an alarm issued for *Streptococcus pneumoniae* has made it possible to reveal a cluster of spectra with close proteinic and temporal distances. This subtree was characterized by the presence of three discriminant peaks (5228.8, 5917.8, and 8974.3 m/z) and the absence of peak 4996.9 m/z. A further investigation on UniProt KB showed that peak 5228.8 is possibly an OxaA protein and that the absent peak may be a transposase.

Conclusions

This example shows how an adequate processing of the bacteria phenotypic expression by using the protein expression coming routinely at low cost by MALDI-TOF mass spectrometry [4,10,11] may show a spectra clustering that support a quasi-real time identification and characterization of clusters suggesting and providing essential information on a potentially epidemic situation. It is a valuable tool for

epidemiological sensemaking and for deciding on the continuation of the epidemiological investigation, in particular the involving of additional costly resources to confirm or invalidate the alarm.

List Of Abbreviations

AP-HM: Assistance Publique – Hôpitaux de Marseille

BALYSES: Bacterial real-time Laboratory-based Surveillance System

BinDA: Binary Discriminant Analysis

BTS: Bacterial Test Standard

IHU-MI: Hospital University Institute Méditerranée infection

MALDI-TOF: Matrix Assisted Laser Desorption Ionization – Time of Flight

MIDaS: Mediterranée Infection Data Warehousing and Surveillance

MSP: Main Spectrum Profiles

REST: representational state transfer

SNR: Signal to Noise Ratio

WGS: whole-genome sequencing

Declarations

Ethics approval and consent to participate

This work relies on our Laboratory Information System, and the surveillance system is a part of its routine statistical exploitation as declared to the French Data Protection Authority (CNIL, declaration number 2139516 v 0). This study, as part of a project granted by the French Ministry of Health for the Hospital Clinical Research Program, has received an ethical approval (2017-024) and has been allowed by the French Data Protection Authority (CNIL decision DR-2018-177). It has also been declared on ClinicalTrials.gov Protocol Registration and Result System (id: NCT03626987).

Consent for publication

Not applicable.

Availability of data and materials

The data from our epidemiological surveillance is part of our routine Laboratory Information System and cannot be directly available on the public domain, but anyone interested in using the data for scientific purpose is free to request permission of specific anonymized extraction from the corresponding author: Hervé Chaudet (herve.chaudet@gmail.com).

Competing interests

All authors report no potential conflicts.

Funding

This study has been supported by a grant from French Ministry of Health for the Hospital Clinical Research Program “SpectraSurv: Identification of Protein Markers of Epidemiological and Clinical Interest by MALDI-TOF” (PHRC 2016_098) and from the OpenHealth Institute.

Authors’ contribution

Conceived and designed the study: AGG, GT and HC. Designed and/or performed experiments: AGG and HC. Analyzed and interpreted data: AGG, GT, PEF, DR and HC. Wrote the manuscript: AGG, GT and HC. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. LANGMUIR AD. The surveillance of communicable diseases of national importance. N Engl J Med. 1963;268:182-192.
2. Thacker SB, Birkhead GS. *Surveillance Field Epidemiology, 2nd* (eds Gregg M) 26-29 (Oxford University Press New York, 2002).
3. Abat C, Chaudet H, Rolain JM, Colson P, Raoult D. Traditional and syndromic surveillance of infectious diseases and pathogens. Int J Infect Dis. 2016;48:22-28.
4. Seng P, Drancourt M, Gouriet F, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis. 2009;49(4):543-551.
5. Chaudet H, Pellegrin L, Gaudin C, Texier G, Queyriaux B, Meynard JB, Boutin JP. A model-Based architecture for supporting situational diagnosis in real- time surveillance systems, *Dis. Surveill*. 2007;4:152.
6. Sintchenko V, Gallego B. Laboratory-guided detection of disease outbreaks: three generations of surveillance systems. Arch Pathol Lab Med. 2009;133(6):916-925.

7. Abat C, Chaudet H, Colson P, Rolain JM, Raoult D. Real-Time Microbiology Laboratory Surveillance System to Detect Abnormal Events and Emerging Infections, Marseille, France. Emerg Infect Dis. 2015;21(8):1302-1310.
8. Foxman B, Riley L. Molecular epidemiology: focus on infection. Am J Epidemiol. 2001;153(12):1135-1141.
9. Sintchenko V, Iredell JR, Gilbert GL. Pathogen profiling for disease management and surveillance. Nat Rev Microbiol. 2007;5(6):464-470.
10. Christner M, Trusch M, Rohde H, et al. Rapid MALDI-TOF mass spectrometry strain typing during a large outbreak of Shiga-Toxigenic Escherichia coli. PLoS One. 2014;9(7):e101924.
11. Griffin PM, Price GR, Schooneveldt JM, et al. Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify vancomycin-resistant enterococci and investigate the epidemiology of an outbreak. J Clin Microbiol. 2012;50(9):2918-2931.
12. Berrazeg M, Diene SM, Drissi M, et al. Biotyping of multidrug-resistant Klebsiella pneumoniae clinical isolates from France and Algeria using MALDI-TOF MS. PLoS One. 2013;8(4):e61428.
13. Khennouchi NC, Loucif L, Boutefnouchet N, Allag H, Rolain JM. MALDI-TOF MS as a Tool To Detect a Nosocomial Outbreak of Extended-Spectrum-β-Lactamase- and ArmA Methyltransferase-Producing Enterobacter cloacae Clinical Isolates in Algeria. Antimicrob Agents Chemother. 2015;59(10):6477-6483.
14. Mlaga KD, Dubourg G, Abat C, et al. Using MALDI-TOF MS typing method to decipher outbreak: the case of Staphylococcus saprophyticus causing urinary tract infections (UTIs) in Marseille, France. Eur J Clin Microbiol Infect Dis. 2017;36(12):2371-2377.
15. Texier G, Pellegrin L, Vignal C, Meynard JB, Deparis X, Chaudet H. Dealing with uncertainty when using a surveillance system. Int J Med Inform. 2017;104:65-73.
16. Salmon M, Schumacher D, Höhle M. Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance. *Journal of Statistical Software*. 2016;**70**(10):1–35.
17. R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
18. Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics. 2012;28(17):2270-2271.
19. Michael H, Hornik K, Buchta C. Getting Things in Order: An Introduction to the R Package seriation. *Journal of Statistical Software* [Online], 2008;25(3):1-34.
20. Gibb S, Strimmer K. Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis. Bioinformatics. 2015;31(19):3156-3162.
21. Palarea-Albaladejo J, Mclean K, Wright F, Smith DGE. MALDIrppa: quality control and robust analysis for mass spectrometry data. Bioinformatics. 2018;34(3):522-523.
22. Giraud-Gatineau A, Texier G, Garnotel E, Raoult D, Chaudet H. Insights Into Subspecies Discrimination Potentiality From Bacteria MALDI-TOF Mass Spectra by Using Data Mining and Diversity Studies.

Front Microbiol. 2020;11:1931.

23. Gruvaeus G, Wainer H. Two additions to hierarchical cluster analysis. *Br J Math Stat Psychol*. 1972;25:200–206.
24. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ. 2015;350:h1314.
25. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. 2010;4:40–79.
26. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1736-1788.
27. Herd M, Kocks C. Gene fragments distinguishing an epidemic-associated strain from a virulent prototype strain of Listeria monocytogenes belong to a distinct functional subset of genes and partially cross-hybridize with other Listeria species. Infect Immun. 2001;69(6):3972-3979.
28. Faruque SM, Chowdhury N, Kamruzzaman M, et al. Genetic diversity and virulence potential of environmental Vibrio cholerae population in a cholera-endemic area. Proc Natl Acad Sci U S A. 2004;101(7):2123-2128.
29. Freitas AR, Tedim AP, Francia MV, et al. Multilevel population genetic analysis of vanA and vanB Enterococcus faecium causing nosocomial outbreaks in 27 countries (1986-2012). J Antimicrob Chemother. 2016;71(12):3351-3366.
30. Bryant JM, Grogono DM, Greaves D, et al. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet. 2013;381(9877):1551-1560.
31. Eyre DW, Cule ML, Wilson DJ, et al. Diverse sources of C. difficile infection identified on whole-genome sequencing. N Engl J Med. 2013;369(13):1195-1205.
32. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Curr Opin Microbiol. 2015;23:62-67.
33. Kan B, Zhou H, Du P, et al. Transforming bacterial disease surveillance and investigation using whole-genome sequence to probe the trace. Front Med. 2018;12(1):23-33.
34. Galardini M, Koumoutsi A, Herrera-Dominguez L, et al. Phenotype inference in an Escherichia coli strain panel. Elife. 2017;6:e31035.

Figures

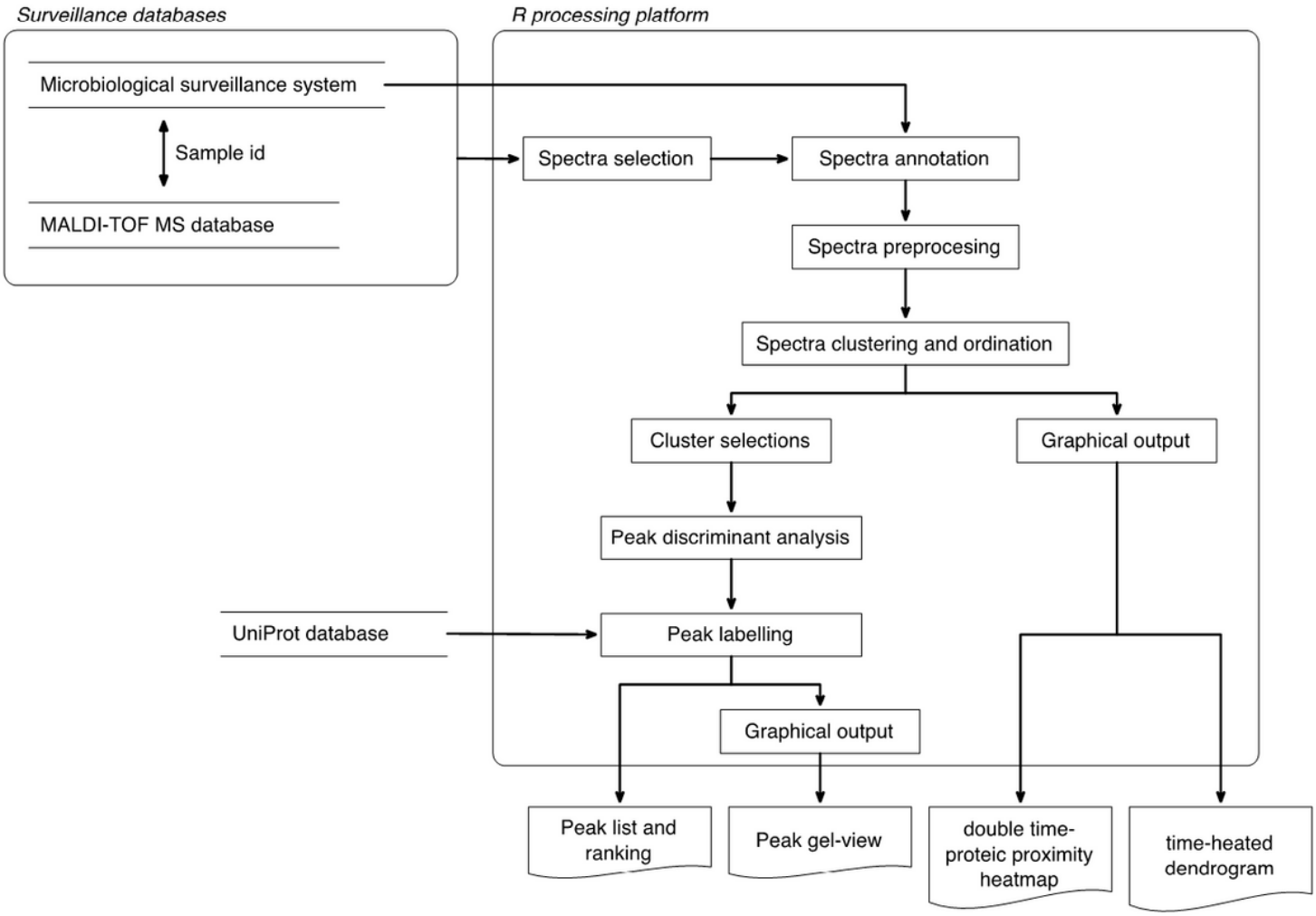

Figure 1

Process flow of the Matrix-Assisted Laser Desorption Ionization Mass Spectrometry mass spectra analysis, from databases to system outputs.

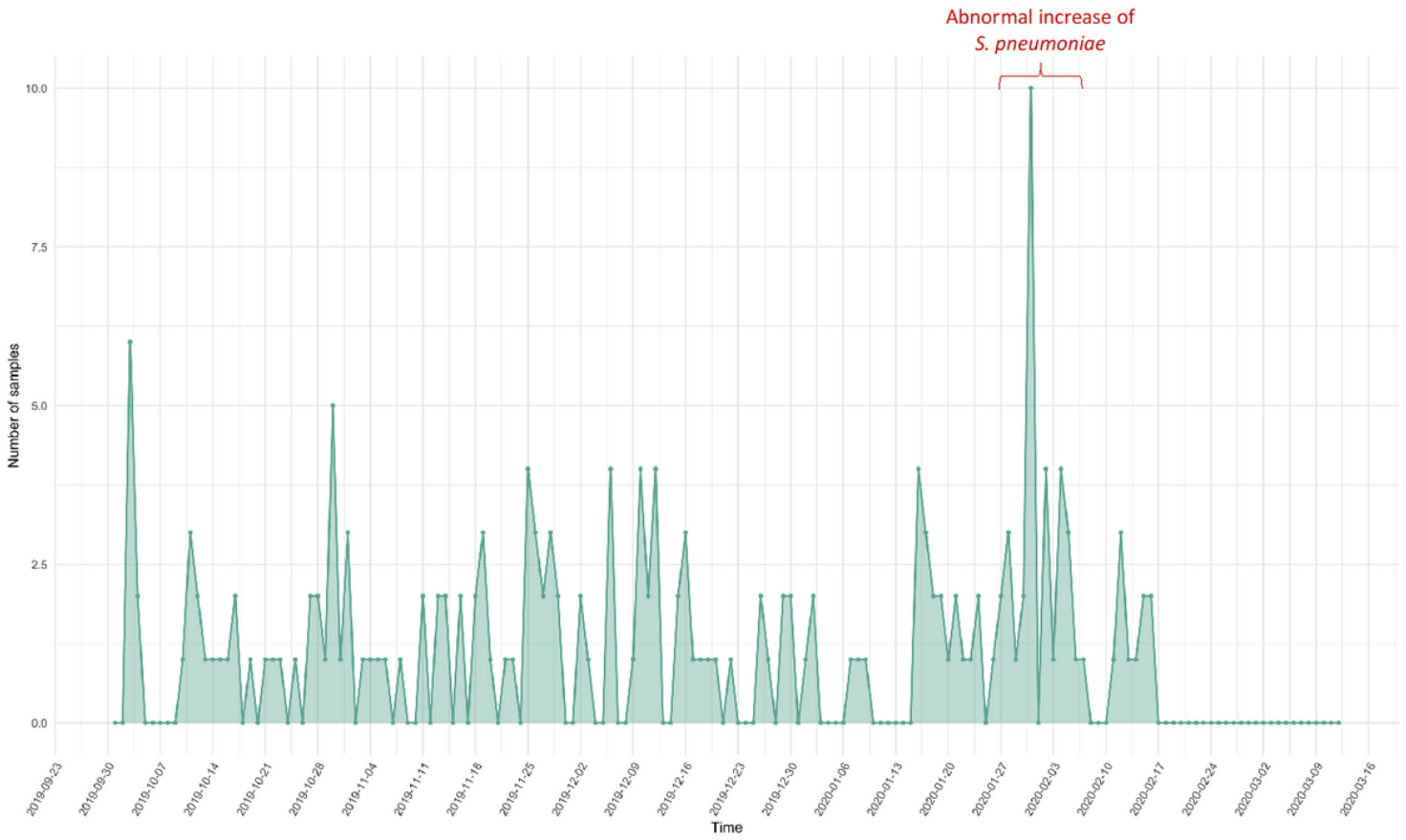

Figure 2

Number of Streptococcus pneumoniae samples from 10 October 2019 to 12 March 2020 at AP-HM, Marseille.

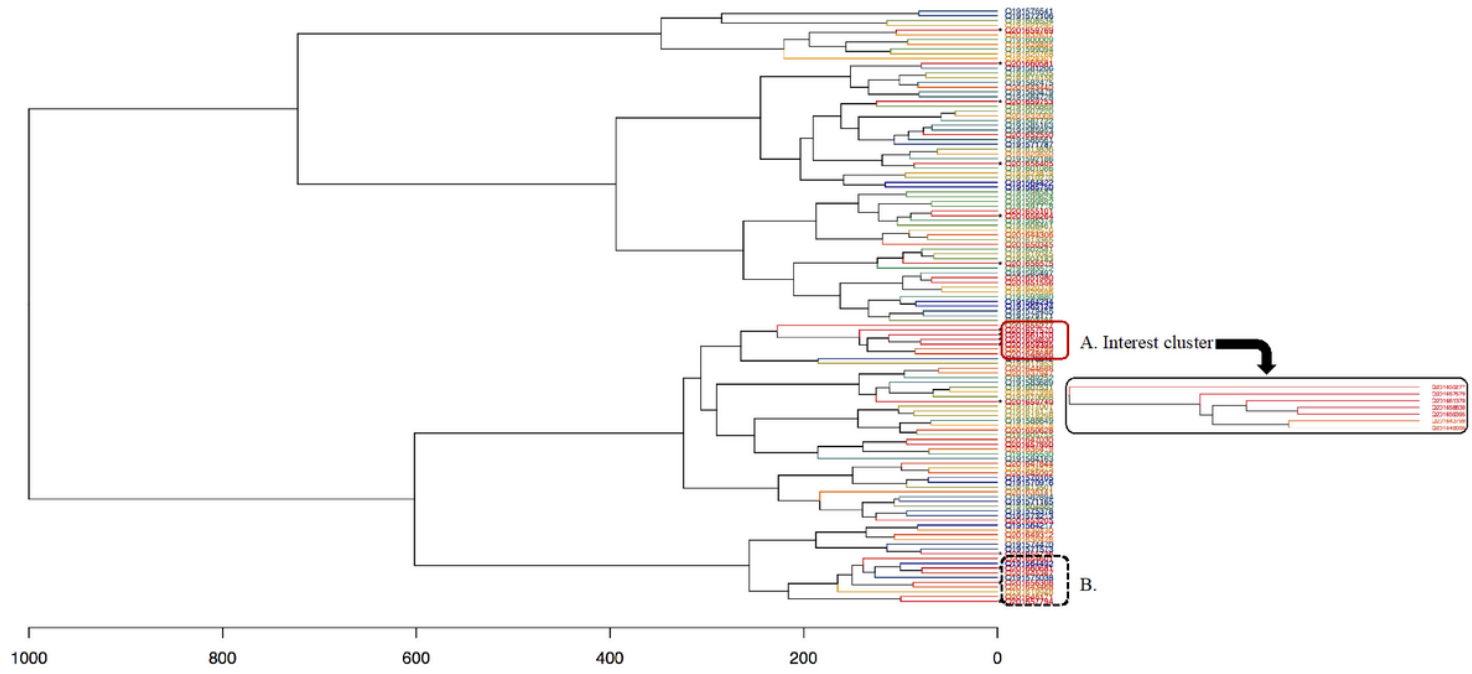

Figure 3

Complete Time-heated dendrogram of the 125 main spectrum profiles of S. pneumoniae illustrating the use of leaf label coloring. The colorscale shows the case recency, from to most ancient (blue color) to the most recent (red color), and then case concomitance. The interest cluster (red square) is indicated with an enlargement of the dendodragrm illustrating the possible leaf labelling using surveillance data. The stars near the leaves are the spectra involved in the alarm emitted by BALYSES.

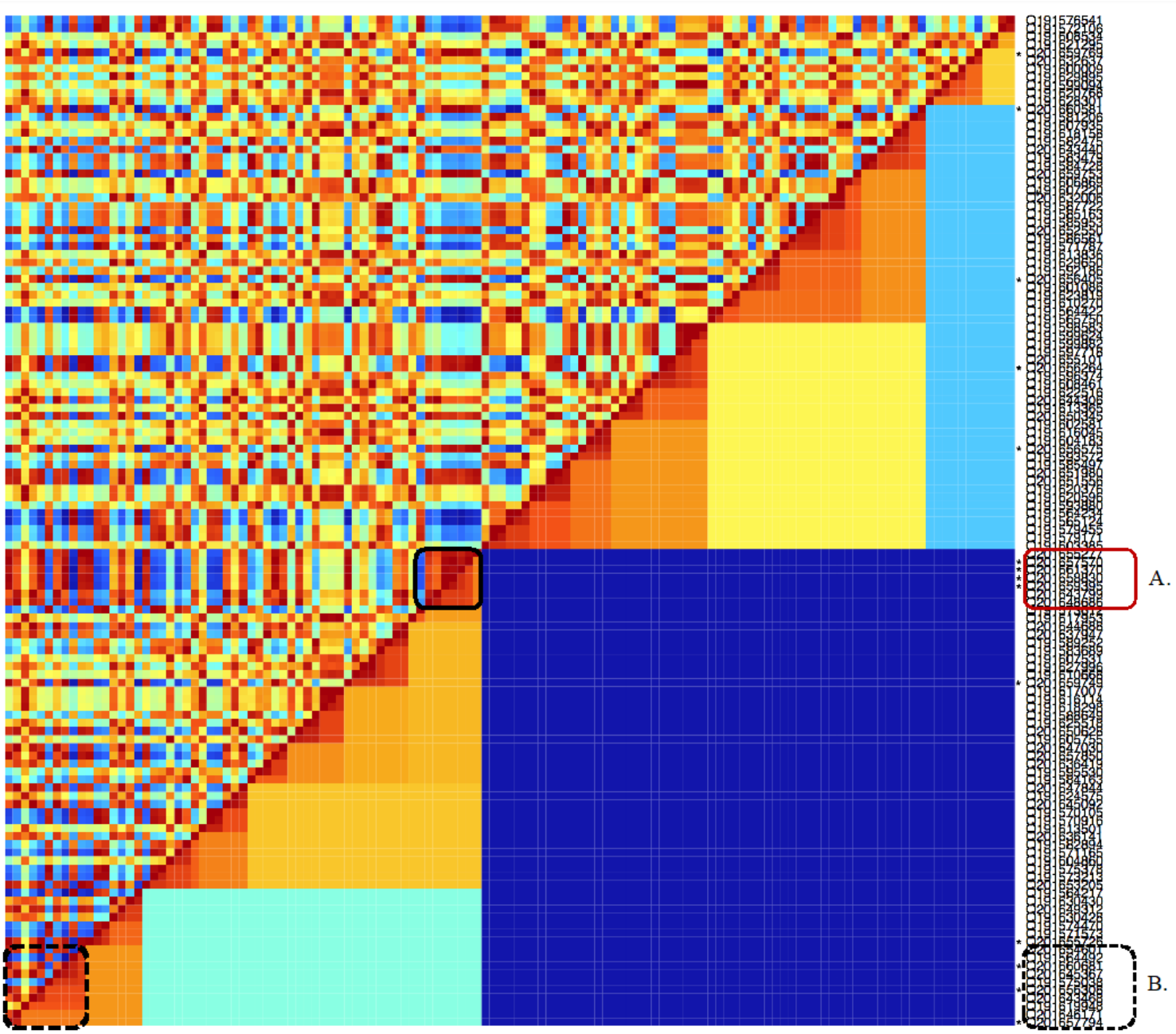

Figure 4

Double time-protein proximity heatmap resulting from the analysis of the 125 main spectrum profiles (MSP) of S. pneumoniae. The interest cluster is indicated. The bottom-right hemi-matrix shows the samples' proteinic proximity. The top-left hemi-matrix shows the samples' time concomitance. The colorscale is the same for the two hemi-matrices: blue corresponds to the largest distances and red to the closest ones. Spectra with closed time-protein distances appears as a square in hot colour along the

matrix diagonal, as subtree A. The stars near the leaves are the spectra involved in the alarm emitted by BALYSES. Subtree D has a less epidemiological interest with a temporal heterogeneity.

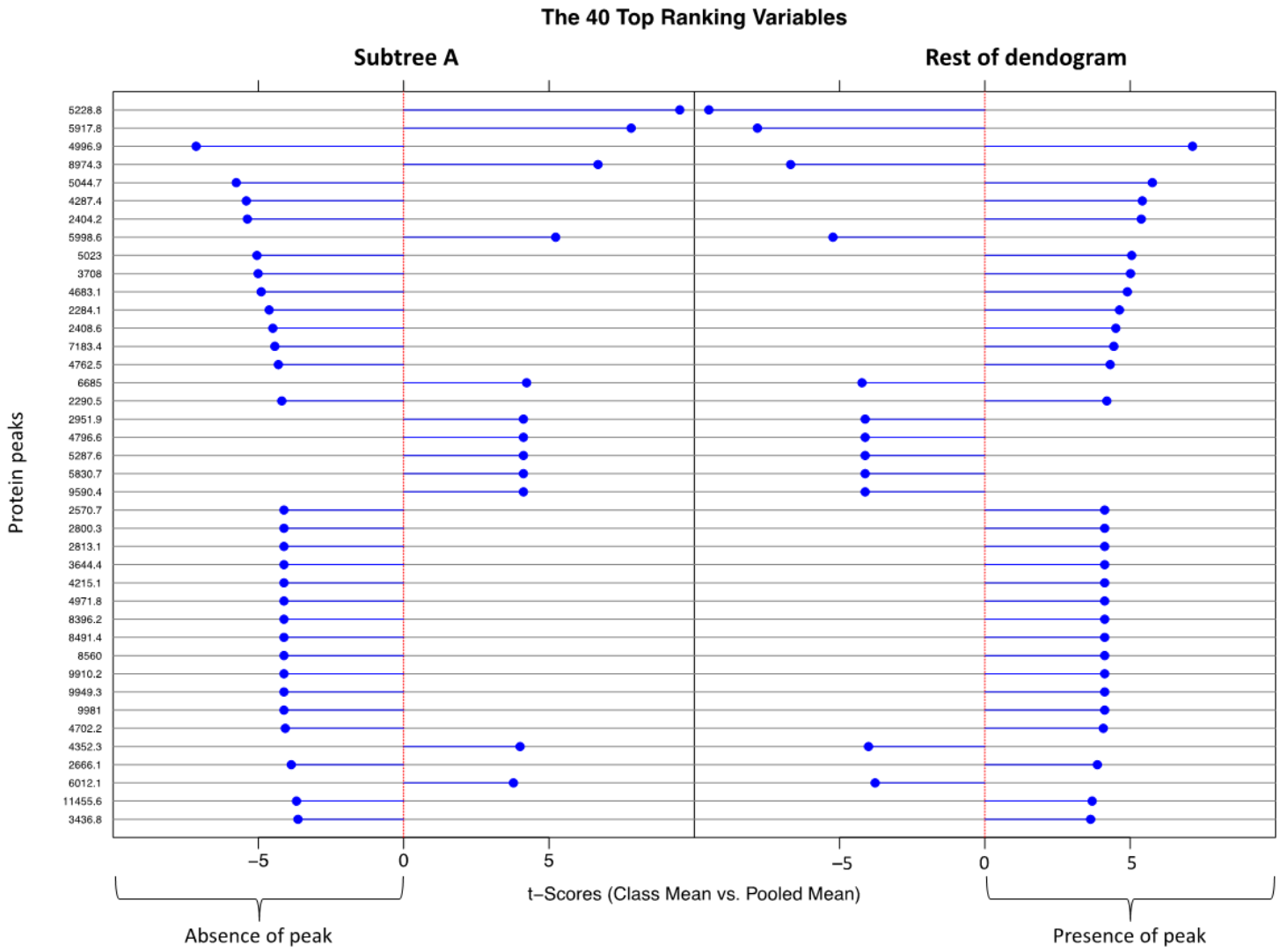

Figure 5

Binary discriminant analysis of the 125 main spectrum profiles (MSP) of S. pneumoniae showing the 40 top ranking peaks contrasting the 7 samples belonging to interest cluster against the other ones. Peaks are indicated using their m/z. For each selected peak the entropic ranking t-score is represented, positive when the peak is associated with the group.