Attenuated total reflectance–Fourier transform infrared spectral profiling of bacteria and yeasts for routine microbial diagnostics

By

Tien My Lisa Lam

Department of Food Science and Agricultural Chemistry

McGill University, Montreal

May 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

SHORT TITLE

ATR-FTIR spectral profiling of microorganisms for routine microbial diagnostics

ABSTRACT

The present work addresses the potential applicability of attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy in routine microbiology for the identification of bacteria and yeasts. The infrared (IR) spectrum of microorganisms acquired directly from initial culture is representative of their biochemical composition and is referred to as a whole-organism fingerprint. However, identifying the species of a microbial isolate by matching its IR spectrum against spectra of reference strains in a spectral database becomes increasingly unreliable as the number of candidate species represented in the database increases. “Spectrotyping” is a novel approach centered on the process of determining the differences between spectra (of microorganisms) based on the absence or presence and relative intensities of particular IR absorption bands and was assessed in the present work for differentiation of microorganisms at the species level. At the outset of this work, a standardized operating protocol was elaborated and tested with diverse microbial strains to ensure that the methodologies developed in the subsequent research would be directly transferable to external sites for validation and potentially routine implementation. Clinical and food isolates of bacteria (*n*=2619) and yeasts (*n*=391) obtained from the frozen collections of 9 microbiology laboratories were cultured in accordance with the standardized operating protocol, and their ATR-FTIR spectra were acquired in triplicate by directly transferring colonies from the culture plate onto the ATR sampling surface. For the spectrotyping of bacteria, averaged triplicate spectra of the 2619 bacterial isolates were divided into training and test sets where isolates in the test set were selected based on stratified random sampling of ~30% of the total isolates in each group (species). A multitude of multivariate spectral analysis methods were investigated in the development and validation of the prediction models. Predictions of Gram-stain type, genus, and species were then obtained from this model for the spectra in the test set. The prediction model yielded 99.3% and 99.4% correct Gram-stain type identification for Gram-negative (*n*=272) and Gram-positive bacteria (*n*=312), respectively. At the genus level, 100% correct identification was achieved for clinically relevant microorganisms. Moreover, at the species level, 92.6% and 99.1% correct identification were obtained for Gram-negative and Gram-positive bacteria, respectively. An external validation study of the ATR-FTIR spectroscopy-based bacteria

identification method employing this prediction model was conducted in a hospital clinical microbiology laboratory over a 3-month period. Among 391 bacterial isolates identified by the ATR-FTIR spectroscopy-based method in parallel with the laboratory's routine identification methods, 363 isolates belonged to species represented in the training set, and 98.7%, 91.3% and 98.4% were correctly identified at the Gram-stain type, genus, and species level, respectively. For yeast identification, a sequential pairwise multitier prediction model was developed with a training set consisting of 261 frozen isolates (previously identified by reference methods) encompassing 12 genera and 65 species. The ATR-FTIR spectroscopy-based method for yeast identification employing this prediction model was ultimately evaluated in a multicenter study (unsupervised) encompassing 6 clinical microbiology laboratories resulting in 98.3% correct species identification with no misidentification of 534 collected isolates. The validation studies of the ATR-FTIR spectroscopy-based methods for bacteria and yeasts based on the prediction models developed in this research demonstrated comparable results to well-accepted matrix-assisted laser desorption ionization–time of flight mass spectrometry, paving the way for future accreditation as a routine tool for microbial identification.

RÉSUMÉ

Le présent travail aborde l'applicabilité potentielle de la spectroscopie infrarouge à réflectance totale atténuée par transformée de Fourier (ATR-FTIR) en microbiologie de routine pour l'identification des micro-organismes. Le spectre infrarouge (IR) des micro-organismes, acquis directement à partir de la culture initiale, est représentatif de leur composition biochimique et est appelé empreinte digitale de tout l'organisme. Cependant, l'identification de l'espèce d'un isolat microbien en faisant correspondre son spectre IR avec des spectres de souches de référence dans une base de données spectrale devient de moins en moins fiable à mesure que le nombre d'espèces candidates représentées dans la base de données augmente. Le « spectrotyping » est une nouvelle approche centrée sur le processus de détermination des différences entre les spectres de micro-organismes en fonction de l'absence ou de la présence et des intensités relatives de bandes d'absorption l'IR particulières et a été évaluée dans le présent travail pour la différenciation des micro-organismes au niveau de l'espèce. Au début de ce travail, un protocole opératoire normalisé (SOP) a été élaboré et testé avec diverses souches microbiennes pour s'assurer que les méthodologies développées dans la recherche ultérieure seraient directement transférables à des sites externes pour validation et éventuellement mise en œuvre de routine. Des isolats cliniques et alimentaires de bactéries (n=2619) et de levures (n=391) obtenus à partir des collections congelées de 9 laboratoires de microbiologie ont été cultivés conformément au SOP, et leurs spectres ATR-FTIR ont été acquis par transfert direct des colonies de la plaque de culture. Pour le spectrotypage bactérien, un échantillonnage aléatoire stratifié d'environ 30 % du total des isolats dans chaque groupe (espèce) des 2619 isolats a été divisé pour un ensemble de tests. De nombreuses de méthodes d'analyse spectrale multivariée ont été étudiées dans le développement et la validation des modèles de prédiction. Les prédictions du type Gram-stain, du genre et des espèces ont ensuite été obtenues. Le modèle de prédiction a donné 99,3 % et 99,4 % d'identification correcte du type Gram-stain pour les bactéries Gram-négatives (n=272) et Gram-positives (n=312), respectivement. Au niveau du genre, une identification correcte à 100 % a été obtenue et au niveau de l'espèce, 92,6 % et 99,1 % d'identification correcte pour les bactéries Gram-négatives et Gram-positives, respectivement. Une étude de validation externe de la méthode d'identification

des bactéries basée sur la spectroscopie ATR-FTIR utilisant ce modèle de prédiction a été menée dans un laboratoire de microbiologie clinique d'un hôpital pendent 3 mois. Parmi 391 isolats bactériens identifiés par la spectroscopie ATR-FTIR en parallèle avec les méthodes d'identification de routine du laboratoire, 363 isolats appartenaient à des espèces représentées dans l'ensemble d'apprentissage, et 98,7 %, 91,3 % et 98,4 % ont été correctement identifiés au type Gram-stain, le genre et l'espèce, respectivement. Pour l'identification des levures, un modèle de prédiction a été développé avec un ensemble d'apprentissage composé de 261 isolats congelés. La méthode basée sur la spectroscopie ATR-FTIR pour l'identification des levures utilisant ce modèle de prédiction a finalement été évaluée dans une étude multicentrique englobant 6 laboratoires de microbiologie clinique, aboutissant à 98,3 % d'identification correcte des espèces sans erreur d'identification de 534 isolats collectés. Les études de validation des méthodes basées sur la spectroscopie ATR-FTIR pour les micro-organismes basés sur les modèles de prédiction développés dans cette recherche ont démontré des résultats comparables aux méthodes d'identification de routine, ouvrant la voie à une future accréditation en tant qu'outil de routine pour l'identification microbienne.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Dr. Ashraf A. Ismail for his immense guidance and support. It wasn't an easy ride pursuing the degree; however, his encouragement, scientific discussions and curiosity had significantly developed my attitude and respect towards scientific research that I will forever carry with me. Behind the scenes, I would like to further extend my thanks to the strong foundation, logical and essential scientific editor of the research group, Dr. Jacqueline A. Sedman. Without her inputs and immaculate editing, the work would not have been accomplished. Additionally, I would like to thank my colleague and good friend Tamao Tsutsumi who has been pursuing her doctoral degree in parallel with me. She has provided me with great conversations, support, motivation, countless late nights in the lab and delicious baked goods. Without her presence, my Ph.D. journey would have been dull. And of course, I would also like to thank my other colleagues (former and current; Dr. Ran Tao, Dr. Mazen Behadi, Xin Di Zhu and Hayline Kim) of the McGill IR Group for all that they have contributed to my journey and all the food we got to enjoy together. Also, a vital component for the completion of my research, I would like to thank Dr. Andrew Ghelter for leaving behind and updating his software to facilitate quicker data analysis.

I would also like to express my deepest appreciation to Dr. Philippe Dufresne at Laboratoire de Santé Publique du Québec for taking me in as a one of his research students. I have acquired invaluable experience under his supervision and without his resources and guidance, the work in the thesis could not have been accomplished.

I would like to give special thanks to McGill University for awarding me the Schulich Fellowship Award from the generous donation from Seymour Schulich.

Lastly, but far from least, I would like to give a huge thank you to my parents, family, and friends for providing me with unlimited unconditional love and support. Without the strong foundation from my family, I would not be where I am standing today or have the mental capacity for completing my degree. I am forever grateful to all that have contributed to my many failures and successes; this journey was surely a group effort and I look forward into starting my next endeavors with all that I have obtained from this process.

CONTRIBUTION TO ORIGINAL KNOWLEDGE

The overall goal of the thesis was to advance the development of ATR-FTIR spectroscopy as a rapid and cost-effective technique for both microbial identification and strain typing of bacteria and yeasts. Extensive research on microbial analysis by FTIR spectroscopy has been reported but limited studies have employed the ATR mode of spectral acquisition. ATR-FTIR spectroscopy was evaluated as part of the work of the thesis to reduce time and cost of microbial analysis. The work of the thesis aimed at the development and evaluation of the ATR-FTIR-based method for microbial identification at the gram-type, genus, species, serotype, and strain-type level for both clinically- and food-relevant microorganisms. The major contributions to knowledge resulting from this work are summarized below.

1. **Standardized microbial sample preparation, spectral acquisition parameters and spectral quality checks for the ATR-FTIR-based microbial identification technique for clinically and food-relevant bacteria and yeasts**

A standard operating procedure was developed to address limitations of the ATR-FTIR-based microbial identification technique such as, sample preparation, polymicrobial culture plate assessment, spectral heterogeneity, and varying moisture content of samples, to generate highly reproducible spectra of microorganisms by ATR-FTIR spectroscopy. This work eliminated major environmental and sample preparation variances between spectra of same-species, thereby generating reliable and consistent ATR-FTIR spectra of intact microbial cells.

2. **Developed and evaluated a multilevel classifier technique with multivariate statistical analysis and machine learning for Gram-stain type, genus, species, and serogroup microbial identification and discrimination**

The thesis's work encompasses the largest ATR-FTIR spectral collection (collected retrospectively and prospectively) of over 7000 well-characterized isolates from both clinical and food microbiology laboratories. Various advanced mathematical classification algorithms were evaluated to produce multivariate discrimination models for the identification and typing of microorganisms solely based on differences in their ATR-FTIR spectra (spectrotyping). Through central (bacteria and yeasts), multicentral (yeasts) and international (yeasts) evaluation, over 95% correct identification was achieved for Gram-

stain type, genus, species, and serogroups in comparison to current identification methods (including carbon assimilation tests, MALDI-TOF MS, molecular-based typing techniques and WGS).

3. **Developed a technique for strain-typing of highly clonal *C. parapsilosis* isolated from a hospital outbreak**

The genome of yeasts is greater than 10 Mb, making whole genome sequencing both costly and time-consuming and requires highly trained laboratory technicians and bioinformatic experts. Microsatellite molecular strain-typing of *C. parapsilosis* lacks sensitivity of WGS and results are not always reproducible. This work has demonstrated the potential of ATR-FTIR spectroscopy as a rapid (1-minute), easily implemented and sensitive method for identifying potential outbreaks and act as a tool for infection control. The ATR-FTIR method offers real-time analysis for surveillance and outbreak control.

4. **Provided multiple feasibility studies to further advance FTIR spectroscopy in clinical and food microbiology**

In addition to the other contribution to knowledge, FTIR spectroscopy was applied to multiple feasibility studies such as ATR-FTIR spectroscopy for mold identification and evaluating FTIR spectroscopy for batch screening, surveillance of patient health status, antimicrobial resistance discrimination and correlating phenotypic, genotypic characteristics to spectrotypic characteristics. It is important to determine the cause of failure and success to truly appreciate and understand the extent of what can be extracted from ATR-FTIR spectra of microorganisms. As such, the work has laid the path for future researchers to advance FTIR spectroscopy for multiple applications of strain typing.

CONTRIBUTION OF AUTHORS

This thesis consists of 8 chapters starting with Chapter 1 as the general introduction of the topics presented in the thesis along with the research objectives. Chapter 2 details a literature review pertaining to changing microbial taxonomy, current methods of microbial identification, the gaps between different microbial identification techniques, Fourier transform infrared (FTIR) spectroscopy for clinical microbiology and current studies and advancements of FTIR-based computational techniques for the identification of microorganisms. Chapter 3 outlines a proposed standard sample preparation, data analysis and identification technique to be used throughout Chapters 4-7 mimicking clinical routine workflow by incorporating standard operating protocols of clinical microbiology laboratories. The spectral data used in Chapter 3 were collected from clinical and food-related microorganisms and were used throughout Chapters 4-7 which were provided by Drs. Émilie Vallières, Simon Lésveque, Philippe Dufresne, Simon Dufresne, Me-Linh Luong, Mathew Cheng, Pierre Lebel, Charles Frennett, George Golding, Catherine Carrillo, Thomas Teklemariam, Jason Gotera, Shawn Lockhart, Michelle Hill, Jean Longtin and Mrs. Irene Iugovaz from their respective institutions. Additionally, Miss Tamao Tsutsumi and Xin Di Zhu (both doctoral students under the supervision of Dr. Ismail) have provided a subset of the data for Chapter 4. Dr. Valentina Cifuentes provided all strains presented in Chapter 7 and Dr. Philippe Dufrene contributed to completing the microsatellite strain typing, purity assessment and assisted in the writing of the results section of this work. Additionally, Dr. Sandrine Moreira performed the whole genome sequencing and analysis of the isolates in Chapter 7.

Chapters 3-7 of this thesis are texts of papers in preparation for publication or have been published as described below. Chapter 3 was written in the format of Methods and Protocols article while Chapters 4 through 7 are research articles. The present author, Miss Lam, was responsible for the experimental design and work (otherwise stated), concepts, data analysis (otherwise stated) and manuscript preparation. Dr. Ashraf A. Ismail is the principal investigator and provided research direction and guidance for the work presented. Dr. Jacqueline A. Sedman acted as the scientific editor and consultant

for all work presented in the thesis. The chapters presented in this thesis were or will be published and/or presented in part(s) at conferences:

CHAPTER 4:

Parts of Chapter 4 were presented at the IDWeek 2019, Washington, DC

Lam, L. M., Ismail, A., Sedman, J., & Vallières, É. (2019, October). 2167. Evaluation of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy for Rapid and Reagent-Free Identification of *Burkholderia* spp. In *Open Forum Infectious Diseases* (Vol. 6).

CHAPTER 5 :

Lam, L. M., Dufresne, P. J., Longtin, J., Sedman, J., & Ismail, A. A. (2019). Reagent-free identification of clinical yeasts by use of attenuated total reflectance Fourier transform infrared spectroscopy. *Journal of clinical microbiology*, *57*(5).

Parts of Chapter 5 were presented at the ASM Microbe 2018 annual conference as a poster and rapid-fire talk (Atlanta, GA, USA)

Lam, L. M., Dufresne, P. J., Longtin, J., Sedman, J., & Ismail, A. A. (2018). Rapid Identification of Routine Clinical *Candida* Species and Discrimination between Fluconazole-Resistant and Fluconazole-Susceptible *Candida auris* by Reagent-Free ATR-FTIR Spectroscopy. American Society for Microbiology: ASM Microbe 2018 Conference. Atlanta, GA.

CHAPTER 6

Lam, L. M., Ismail, A. A., Lévesque, S., Cheng, M. P., Dufresne, S. F., Vallières, É., Luong, M., Sedman, J., Dufresne, P. J. Multicenter evaluation of attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy-based method for rapid identification of clinically relevant yeasts (manuscript accepted with revision August 2021 to the Journal of clinical microbiology)

CHAPTER 7

Parts of Chapter 6 were presented at the ASM Microbe 2019 annual conference as a poster (San Francisco, CA, USA)

Lam, L. M., Dufresne, P. Cifuentes, V., J., Sedman, J., & Ismail, A. A. (2019). Investigation of Suspected *Candida parapsilosis* Hospital Outbreaks Employing Strain Typing by Attenuated Total Reflectance–Fourier Transform Infrared Spectroscopy in Comparison to Microsatellite Genotypic Profiles. American Society for Microbiology: ASM Microbe 2019 Conference. San Francisco, CA.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBRIVIATIONS

ABC	ATP-binding cassette
ANN	Artificial neural networks
AST	Antimicrobial susceptibility testing
ATR	Attenuated total reflectance
Aw	Water activity
BAP	Blood agar
BCC	*Burkholderia cepacia* complex
CDC	Center of Disease Control and Prevention
CFIA	Canadian Food Inspection Agency
CGD	Candida Genome Database
CHRM	Chrome agar
CHUS	Centre hospitalier Sherbrooke
CHUSJ	Centre hospitalier universitaire Sainte-Justine
CLSI	Clinical and Laboratory Standards Institute
CoNS	Coagulase-negative staphylococci
CS	Classification score
CVA	Canonical variate analysis
DA	Discriminant analysis
EPS	Extracellular polymeric substance
ESBLs	Extended-spectrum beta-lactamases
FLU	Fluconazole
FTIR	Fourier transform infrared spectroscopy
GN	Gram-negative
GP	Gram-positive
HCA	Health Canada
HCA	Hierarchical cluster analysis
HR-MAS	High-resolution magic angle spinning
IFIs	Invasive fungal infections
IMA	Inhibitory mold agar
IR	Infrared
ITS	Internal transcribed spacer

k-NN	k-nearest neighbors algorithm
LPS	Lipopolysaccharide
LSPQ	Laboratoire de Santé Publique du Québec
m/z	Mass to charge ratios
MAC	MacConkey agar
MAID	Maidstone Hospital
MALDI-TOF	Matrix-assisted laser desorption/ionization time of flight
MDR	Multi-drug resistance
MIC	Minimum inhibitory concentration
MLC	MultiLevel Classifier
MLST	Multilocus sequence typing
MMA	Miller-Mallinson agar
MRSA	Methicillin resistant *Staphylococcus aureus*
MS	Mass spectrometry
MSMT	Microsatellite molecular typing
MSSA	Methicillin-sensitive *Staphylococcus aureus*
MUHC	McGill University Heath Center
NAC	Non-albicans Candida
NFGNB	Non-fermenting Gram-negative bacilli
NML	National Microbiology Laboratory
NMR	Nuclear magnetic resonance
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDA	Potato Dextrose Agar
PFGE	Pulsed-field gel electrophoresis
PLS	Partial least squares
PMF	Peptide mass fingerprint
QIMR	Queensland Institute of Medical Research
RAPD	Random amplified polymorphic DNA
rDNA	Ribosomal rDNA
RT	Room temperature

SAB	Sabouraud dextrose agar
SABE	Sabouraud Dextrose Agar Emmons
SD	Standard deviation
SID	Simpson’s index of diversity
SIMCA	Soft Independent Modeling of Class Analogy
SNR	Signal-to-noise ratio
SOP	Standard operating procedures
SQ	Spectral quality
SVM	Support vector machines
TF	Transcription factor
TY	Tryptone yeast extract agar
VRE	Vancomycin-resistant enterococci
VSE	Vancomycin-sensitive enterococci
WGS	Whole genome sequencing
WT	Wild type
XLD	Xylose lysine deoxycholate agar

CHAPTER 1. INTRODUCTION

1.1. General introduction

Microorganisms causing infections and illnesses are acquired from hospitals, contaminated foods, person-to-person transmission and can also be acquired from the environment. *Staphylococcus aureus*, *Enterococcus* species, *Escherichia coli*, *Shigella* species, *Candida albicans*, and more recently *Candida auris* are examples of microorganisms that cause infections in both healthy and immunocompromised individuals. Conventionally, for screening and identification, microorganisms are grown on culture agar media and broths. These methods are based on monitoring the biochemical and metabolic profiles of microorganisms during their growth cycle over a 16-48 h period. Although sensitive and inexpensive, conventional methods are time consuming and are not always conclusive – requiring the use of additional identification methods for achieving conclusive results. Precision, accuracy, rapidness, and cost are major considerations for the development of new microbial identification methods. With the advancement in microbial identification tools, immunologically based methods have been developed that are more rapid than conventional techniques. Limitations of immunological-based methods include lower specificity, generation of antibodies to select antigens specific to species and are available to a limited selection of species. Furthermore, identification sensitivity varies between genera and species, and therefore various factors should be considered when developing new methods for species identification. In addition, some microorganisms share genotypic and phenotypic attributes, contributing to the difficulties in species identification. For instance, *E. coli* and *Shigella* spp. are phylogenetically similar but are phenotypically dissimilar where *Shigella* spp. are non-lactose fermenting whereas *E. coli* are lactose fermenters (rarely non). Conventional biochemical identification methods heavily rely on specific characteristics of microorganisms and therefore may lack sensitivity for species with overlapping characteristics. Molecular-based methods [such as polymerase chain reaction (PCR) and whole genome sequencing (WGS)] are more precise for species identification and are used to identify the presence of the antimicrobial resistance gene(s). While WGS is becoming the new gold standard for microbial identification and typing, it is expensive,

time consuming and requires highly trained personnel, making it currently unsuitable for routine use in microbiology laboratories.

The past decade witnessed the emergence and global adoption of new microbial identification techniques based on the use of matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometry (MS). MALDI-TOF MS provides rapid, accurate and cost-efficient microbial identification compared to current molecular methods. However, MALDI-TOF MS requires a high initial investment cost, making it unsuitable for small clinical microbiology laboratories and the food industry (quality control and assurance).

The adoption of MALDI-TOF MS as part of the workflow of a clinical microbiology laboratory opened doors to spectroscopy-based methods for clinical *in vitro* diagnostics such as Raman and Fourier transform infrared (FTIR) spectroscopy. The infrared spectrum of a microorganism can be acquired directly from intact cells (referred to as a whole-organism fingerprint) taken from a culture plate and is representative of their biochemical composition (i.e., lipids, proteins, carbohydrates, etc.), and as such is considered as the most characteristic physical property of the microorganism. Like MALDI-TOF MS, microbial identification employing FTIR spectroscopy is based on spectral fingerprint recognition by comparison of the FTIR spectrum of an unknown isolate with FTIR spectra in a reference library of well characterized microbial strains. While most of the published research to date has involved FTIR spectral acquisition in the transmission mode, which requires several sample processing steps prior to spectral acquisition, the use of attenuated total reflectance (ATR) mode of FTIR spectral acquisition of microorganisms directly acquired from the culture media is more practical for implementation for routine laboratory use. Furthermore, an ATR-FTIR-based spectroscopic method does not require the use of any reagents making it more cost-effective to resource-limited laboratories. The current research will evaluate the specificity and sensitivity of ATR-FTIR spectroscopy for identification and typing of selected microorganisms in comparison to other current identification techniques.

1.2. Research objectives

The research objective of the thesis entails the examination of the applicability of ATR-FTIR spectroscopy for identifying clinical and food relevant microorganisms, and to advance the development of FTIR spectroscopy as a rapid and cost-effective technique for strain-typing of outbreaks.

The specific research objectives are as follows:

i. To develop a standardized protocol for ATR-FTIR spectral acquisition for microbial identification for selective (clinically and food relevant) non-fastidious, aerobic, and facultative anaerobic bacteria and yeasts. Standardization of the ATR-FTIR-based microbial identification method entails defining culturing conditions, spectral acquisition parameters (e.g., number and frequency of background scans, number of sample scans and spectral resolution), spectral processing and pre-processing, and to develop a multilevel classification spectral search technique that will further advance ATR-FTIR spectroscopy as a tool for microorganism identification.

ii. To define, investigate and explore spectrotyping and support vector machine algorithms as a method of classifying selective (clinically and food relevant) non-fastidious, aerobic, and facultative anaerobic microorganisms based on their ATR-FTIR spectral fingerprints.

iii. To use information observed in the ATR-FTIR spectral fingerprints of microorganisms to develop a microorganism identification tool by creating a multitier (based on pairs or multiple groups of spectra) spectral search database based on their spectrotypes of well-characterized bacteria and yeasts.

iv. To evaluate the potential of ATR-FTIR spectroscopy-based microorganism identification for routine species identification of food and clinically relevant species in prediction accuracy at the genus, species, and serogroup level; retrospectively and prospectively identify microorganisms based on differences in their infrared spectra.

v. To multicentrally evaluate the intra- and inter- laboratory data transferability of the ATR-FTIR spectroscopy-based microorganism identification method to advance validation studies for potential implementation.

vi. To compare the efficacy of microbial strain typing by ATR-FTIR spectroscopy in reference to gold standard genotyping methods such as WGS and PCR genetic sequencing for outbreak detection of yeasts to explore and evaluate the ATR-FTIR-based strain typing method.

vii. Lastly, the listed objectives aim at advancing research for ATR-FTIR spectroscopy in clinical and food microbiology, however, there are multiple gaps in which the final objective of the thesis is to address prospective future work to further advance the current field of research.

CHAPTER 2.LITERATURE REVIEW

2.1. Overview of microbial classification of clinical and food-related microorganisms

Since the 1680s, scientists were interested in classifying organisms based on similarities observed physically under the microscope. The work was originally part of Anthonie Van Leeuwenhoek's discovery on living microorganisms, subsequent to developing microscopes (1). With the discovery of microorganisms, biologists naturally navigated towards classification of all newly discovered strains. Classification was originally achieved by categorizing bacteria based on morphology observed under the microscope and was later advanced to lipid extraction and analysis by gas chromatography to group differences in lipid profiles as new species (2). As such, microorganisms were classified by phenotype (genotypic expression that can be observed in various environmental conditions) – with no linkage to previous ancestors. In the recent past decades, many scientists explored classification methods based on varying characteristics of microorganisms, (e.g., colorimetric testing, pulsed-field gel electrophoresis (PFGE), electron microscopy, x-ray diffraction, mass spectrometry (MS), DNA base sequence, rRNA protein profiles and more recently by whole genome sequencing (WGS) (3-6). Through phenotypic and genotypic information, microbiologists were able to develop various classification methods and harvest data to determine relatedness between genera and species.

Although classification of microorganisms is yet to be definitive, WGS is emerging as a standard method for determining phylogeny of microorganisms - resulting in reclassification of existing species to different genera or reclassification of existing genera to other genera [e.g., *Pantoea* genus is now considered to be *Enterobacter* and *Enterococcus* was separated from the *Streptococcus* genus and are classified as its own genus (7-9)].

2.2. Overview of classification of microorganisms

Based on phenotypic classification, for routine identification, bacteria are categorized by their ability to retain crystal violet dye in the cell wall during Gram-staining. Gram-negative (GN) bacteria are microorganisms that do not retain the dye as a result of

having a thin peptidoglycan layer and an inner cell membrane. Within the GN group, there are serval families of clinical and food relevant microorganisms, of which the most significant family is the *Enterobacteriaceae* family. These microorganisms are found naturally in the human and animal microflora and can be isolated from the environment; not all are pathogenic. *Enterobacteriaceae* are bacilli (rod-shaped) or cocci (round), they are mainly motile, are non-spore forming and produce lactic acid by fermenting sugars. Within the *Enterobacteriaceae* family, *Escherichia coli*, *Klebsiella* spp. and *Enterobacter* spp., account for 80-95% of all GN bacteria isolated in clinical practice (10, 11). They are responsible for hospital acquired infections and causes respiratory tract, urinary tract, and bloodstream infections (7). Of food relevance, *Citrobacter* spp., *Salmonella enterica* and *Shigella* spp. are also part of the *Enterobacteriaceae* family and can also cause human infections.

Gram-positive (GP) bacteria retain the crystal violet dye due to the absence of an outer membrane (compared to GN bacteria) where the dye is trapped below the thick peptidoglycan layer. GP bacteria such as staphylococci are found commensally living on the surface of human skin and in the environment. Staphylococci can be opportunistic pathogens and are commonly associated with bacteremia and have a high prevalence of antibiotic resistance (e.g., methicillin-resistant *Staphylococcus aureus,* MRSA) (12). Two other GP bacteria genera of importance include *Listeria* (a major concern in the food industry) and vancomycin-sensitive enterococci and vancomycin-resistant enterococci (VRE) which are prevalent in a clinical setting.

Similar to bacteria, yeasts were once originally classified by their ability to mate (genetic similarity), their physical characteristics (morphology) and differences in metabolic activity. More recently, microbiologists are classifying yeasts by phylogenetics which is reliant on genetic evolution and ancestry. Yeasts can be opportunistic pathogens that are found naturally on our skin, body and contributes to good health. However, yeast infections are the third leading cause of serious bloodstream infection to potentially cause death. Yeasts belonging to *Candida*, *Trichosporon* and *Cryptococcus* genera are known to cause majority of yeast infections. Most common yeast infections come from *Candida* spp. which encompasses approximately 200 species that are expected to be reclassified

into new genera with the advancement of classification tools like WGS. *C. albicans* is associated with approximately 50% of all yeast infections followed by *C*. *glabrata*. Other clinically relevant *Candida* species includes *C. auris, C. dubliniensis*, *C. guilliermondii*, *C. haemulonii*, *C. kefyr*, *C. krusei*, *C. lusitaniae*, *C. parapsilosis* and *C. utilis*. Some *Candida* spp. are morphologically identical but are phylogenetically closely related and are near-distinguishable. An example of such phenomenon is between *C. bracarensis*, *C. nivariensis* and *C. glabrata* which are phenotypically indistinguishable but can also be phylogenetically distinguished. Species identification of phylogenetically related *C. albicans* and *C. dubliniensis* is important due to *C. dubliniensis* gaining resistance to the antifungal, fluconazole (13). *C. glabrata* and *C. krusei* species identification is needed as they are both resistant to fluconazole and fluconazole is a front-line therapy drug (14).

Although speciation is not necessary for all species for appropriate treatment and care, identification of potential pathogenic microorganisms by genera and/or antimicrobial resistance profile is sufficient for rapid screening and will be cost and time efficient. If speciation is necessary, additional methods are available for species identification.

2.3. Current methods for microorganism identification

2.3.1. Biochemical assimilation techniques for microbial identification

Conventional microorganism identification methods include colony morphology, coagulase production, agglutination assays and biochemical tests (fermentation, oxidation, degradation, and hydrolysis of various substrates) (7, 9, 12). Biochemical tests are commercially available and are usually automated to reduce the cost and time of each test. Phenotypic classification is based on morphology, cell structure, cellular metabolism and differences in cellular components such as proteins, DNA, fatty acids, pigments, quinones and surface antigens (15).

Commercially available phenotypic classification techniques utilize the assimilation of various organic compounds such as selected sugars as a carbon-source for microbial growth. Various carbon sources are added to culture media with varying pH that favors growth of specific microorganisms (16). Incubation temperatures, culture media composition and time under aerobic and anaerobic conditions are modulated to inhibit the

growth of certain organisms while favoring the growth of others. For example, species belonging to the *Enterobacteriaceae* family are tentatively identified by the production of indole, ornithine decarboxylase, malonate and acids from sucrose, dulcitol, melibiose and adonitol under defined growth conditions. The *Enterobacteriaceae* family is considered the most prominent group of microorganisms to cause GN bacteria-related infections. Species identification of *Enterobacteriaceae* is vital, as certain species have developed resistance to the following antibiotics: extended-spectrum beta-lactamases, cephalosporinases and/or carbapenems.

Traditionally, Gram-staining and morphology observations can be used to screen the samples, and the use of selective media such as MacConkey agar (MAC) can improve the accuracy of identifying *Enterobacteriaceae* species. In addition to biochemical reaction-based methods, serotyping is another method to phenotypically identify microorganisms (e.g., for *E. coli*, *S. enterica*, *Streptococcus* spp. and *Staphylococcus* spp.). Microscopic characteristics (commonly for yeasts), macroscopic characteristics (morphology) and lack of coagulase production are among other commonly used phenotyping techniques (14). Detailed accepted phenotypic techniques can be found in the global guidelines through the Clinical and Laboratory Standards Institute (17).

Conventionally, biochemical techniques such as multiple carbon assimilation and enzymatic tests are utilized in routine for microbial identification (18, 19). Although different species share similar biochemical test results, the pattern of the results are not identical, as the species do not share identical intracellular enzymes and therefore will not have the same metabolic pathways (20). The difference in biochemical test results of the different species will result in different “biochemical fingerprints” that are unique to identify/discriminate over/between 300 microorganisms of clinical, food and environmental relevance using commercially available test kits or are performed manually (21).

To date, there are hundreds of biochemical tests available, but only a combination of biochemical tests is used to speciate unknown microorganisms (22-26). Currently, VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile) is an automated commercially available multiple biochemical testing system based on enzymatic activity, acidification, carbon

assimilation and whether certain biochemical components are present or not (21). For example, from the three available test cards for microbial identification available for the system (GN bacteria, GP bacteria, and yeasts), the GN bacteria card performs 47 different biochemical tests in microdilution wells for the identification of approximately 150 GN species. Each microwell is filled with 0.002-0.300 mg of substrates (some chromogenic and others not) where a 3 mL 0.50-0.63 McFarland suspension of the microorganism (and sometimes filled with a color indicator such as bromocresol purple, phenol red or methyl red) is automatically distributed to the microdilution wells where it is incubated to facilitate the biochemical reactions. Colorimetric measurements are acquired every 15 minutes until completion; in approximately 10 hours or less to obtain a biochemical profile (set of positive and negative results) to match to known biochemical profiles in a database for the identification of the microorganism (21, 24).

Comparing the performance of VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile) to molecular and other phenotypic identification methods, VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile) resulted in higher accordance rates (+4%) compared to other phenotypic methods than genotypic methods (Table 2.1). Slow metabolic rates and high phenotypic diversity of certain species such as coagulase-negative staphylococci (CoNS) can result in lower rates of correct identification at the species-level (27). *Enterobacteriaceae* are GN bacteria that are commonly routinely isolated and requires species identification for appropriate treatment, however, correct species identification ranges between 48-89% relative to molecular methods. The low performance of the technique for GN bacteria species identification requires the use of additional methods to obtain conclusive results.

In addition to automated biochemical testing for species identification using the automated VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile) system, it also has the ability to perform antimicrobial susceptibility testing (AST). AST is based on the assimilation of various organic compounds such as selected sugars as a carbon-source for microbial growth in the presence of target antimicrobials. These tests are routinely carried out for *Burkholderia cepacia* complex (BCC) species (resistant to aminoglycoside and polymyxin), MRSA, VRE and *C. krusei*. AST is normally carried out after species identification (14).

Table 2.1. General summary of meta-analyses for the performance of VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile), an automated microbial biochemical reaction-based phenotypic identification technique

	Comparative method	Correct identification in %
Gram-stain type		
Negative	Gram stain	94.37–99.04
Positive	Gram stain	95.48–99.72
Genus		
Gram negative	Molecular	74.85–98.84
Gram positive	Molecular	84.09–99.48
Species		
Gram negative	Molecular	48.06–89.23
	Conventional	92.74–97.70
Gram positive	Molecular	72.65–91.14
	Conventional	94.06–98.11
Yeasts	Molecular	83.0-97.0
	Conventional	89.0-98.0
References	(28-33)	

2.3.2. Matrix assisted laser desorption ionization time of flight mass spectrometry as a tool for microorganism identification

Over the past decade, the use of phenotypic methods has been replaced or supplemented by the application of matrix assisted laser desorption ionization time of flight (MALDI-TOF) MS for the identification of microorganisms. Briefly, microbial samples are smeared over a small area (3-6 mm in diameter) on a target MALDI-TOF MS plate and allowed to dry. A small aliquot (~1 µL) of an organic solvent containing α-cyano-4-hydroxycinnamic acid is added to the sample and is also allowed to dry on the benchtop. The target plate is placed under high vacuum in the mass spectrometer and the sample is irradiated with a high energy nitrogen laser light. Ionization of the organic matrix results in the generation of high energy fragments of α-cyano-4-hydroxycinnamic acid that acts as an intermediary for energy absorption (Figure 2.1). Subsequently, the energy absorption leads to the rupture of the microorganism, then releasing charged proteins which are accelerated under vacuum through a fixed potential - allowing for the separation of ions according to their mass to charge ratios. These ions are then analyzed using a TOF analyzer, which measures the duration of time an ion takes to travel the length of the flight tube of the mass spectrometer (Figure 2.1). Data derived from the TOF analyzer

is used to form a peptide mass fingerprint (PMF) spectrum. In the case of microbial identification, the spectrometer uses the mass ranges from 2-20 kDa which represents 60-70% of all ribosomal proteins in the dry weight of the microorganism. The PMF spectrum (acquired in ~60-90 seconds per sample) of the unknown sample can then be compared to known PMF spectra in a database (34). A reference microorganism (e.g., an ATCC *E. coli* strain) is used to verify the performance of the system.

Comparing with conventional phenotypic microorganism identification techniques, which takes 8-16 hours to obtain a result after initial culture, many researchers are favoring the use of MALDI-TOF MS for its speed of analysis (~3 minutes per sample in duplicates) and lower cost-per-test. There is however, a very high initial capital cost of the equipment which may not be suitable for resource-limited microbiology laboratories.

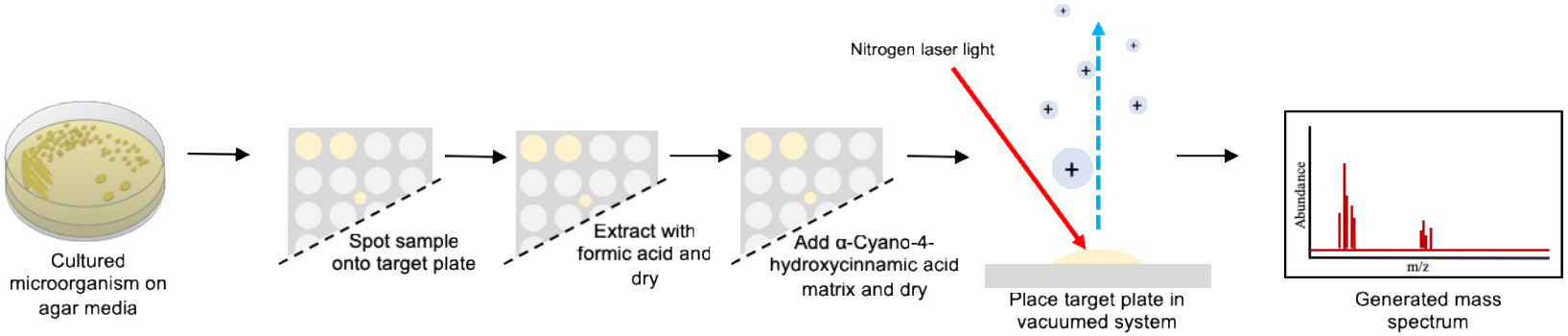


Figure 2.1. Standard laboratory workflow for MALDI-TOF MS-based microbial identification

2.4. Current strain typing methods for bacteria and yeast

Microorganism strain typing is typically performed to gather information on the phylogenetic descendants (microbial evolution) for use in epidemiological investigations (35). From the 1930s until the 1980s, phenotyping was widely used (serotyping, phage typing, biotyping and colicin typing). Subsequently, genomic-based typing methods such as PFGE, rep-Polymerase chain reaction (PCR), multilocus sequence typing (MLST) and WGS became more widespread. Attempts to develop infrared spectroscopy-based strain typing methods were first reported in 1988 (36, 37) and is slowly gaining momentum (Table 2.5, Table 2.6, Table 2.7, Table 2.8 and Table 2.9).

A brief overview of the advantages and disadvantages of selected genomic-based strain typing methods are presented in the following subheadings:

2.4.1. Pulsed field gel electrophoresis

PFGE is the “gold standard” for many public health agencies such as Center of Disease Control and Prevention (CDC) and Health Canada (38). A reported disadvantage of PFGE is associated with the use of non-standardized methods which leads to low inter-laboratory reproducibility (39, 40). PulseNet was created to standardize protocols for subtyping foodborne pathogens, where participating laboratories results are documented and placed in a global database which can be accessed globally and referenced to. PFGE requires extensive training, use of reagents and is highly laborious. The method is suitable for detecting and preventing foodborne outbreaks on a national level, but, in a hospital setting, it is too costly and time consuming. It takes approximately 24 hours for the full analysis and up to 3 or 4 days for robust results (41).

2.4.2. Polymerase chain reaction-based identification technique

PCR is the amplification of DNA fragments subsequent to the treatment with specific enzymes that cleave DNA at designated sites followed by the isolation of the DNA fragments using a primer sequence and amplification by a DNA polymerase (42). While PCR-based methods proved to be sensitive and accurate, PCR requires reagents that are specific to the gene of interest for amplification, and it is also time-consuming, taking

roughly 6 to 48 hours for complete analysis (43-46). For microbial identification, PCR is used to determine whether a target gene (species-specific or an antimicrobial resistance gene) is present or not. For non-targeted identification applications, PCR is coupled with 16S rRNA/rDNA gene sequencing to identify a limited number of species by matching the gene sequence obtained with those from a 16S rRNA/rDNA gene sequence database of microorganisms (43).

Moreover, with advancement of PCR technology, MLST was developed for detecting (the presence of/or absence of) antimicrobial resistant genes and identifying bacterial clones (47-50). These advancements did not completely close the gap between high discriminatory power between all relevant species and the detection of important phenotypic variation of the cell surface resulting from genetic changes (35). WGS is increasingly applied to strain-typing and efforts are underway to elucidate how changes in the genome of a microorganism affect its phenotypic attributes using bioinformatics (51, 52).

2.4.3. MALDI-TOF MS for strain typing of microorganisms

The increasing prevalence of MALDI-TOF MS instruments for microbial identification has led researchers to evaluate its potential use for strain-typing, but the commercially available MALDI-TOF MS spectra are based on the use of a limited mass range of proteins (mainly ribosomal proteins), may not be suitable for strain typing (34). Furthermore, there currently is a lack of method standardization for strain typing analysis, yielding results that are inconsistent relative to other bacterial typing techniques (e.g., *Salmonella* serovars and clonal *Klebsiella oxytoca* and *Acinetobacter baumannii*) (53-55). Despite these limitations, some studies have reported success for selected species of GN and GP bacteria (56).

2.5. Potential use of Fourier transform infrared spectroscopy for microorganism identification and strain-typing

2.5.1. Development overview of Fourier transform infrared spectroscopy-based techniques for microorganism identification and strain-typing

Fourier transform infrared (FTIR) spectroscopy is a well-established technology that has been highly researched in various fields of science including chemistry, forensics, environmental sciences, pharmaceuticals and food science, and in the recent decades, microbiology (57-59). FTIR spectroscopy probes molecular vibrations of chemical bonds possessing a dipole moment. These bonds vibrate at a specific vibrational frequency which can be observed in the infrared spectral region of the electromagnetic spectrum (60). To acquire an infrared spectrum, the infrared beam generated by heating a ceramic element passes through an interferometer producing a modulated infrared beam. A sample placed in the path of the modulated beam absorbs selected wavelengths (dependent on the chemical or biochemical composition of the sample) and the remaining unabsorbed energy impinges on an infrared detector. The absorption frequencies are identified by ratioing the spectrum of the sample by the spectrum of the modulated infrared beam in the absence of the sample.

In microbiology, FTIR spectroscopy has been employed in the study of examining microbial biofilms such as extracellular polymeric substances (EPS). In a study, over a 20-day period, FTIR spectra of live *Pseudomonas* spp. were acquired, and the biofilm formation of the bacteria was physically and spectrally observed. FTIR absorption bands were observed between 1200-900 cm^{-1} were assigned to EPS (61). Moreover, cell wall structural changes may also be examined by FTIR spectroscopy and can be useful in the study of drug resistance mechanism of the bacterium (62). Interestingly, the changes in the chemical composition of lipopolysaccharides (LPS) between different species (e.g., changes in the number of glucosamine disaccharide with six or seven acyl chains) can be used to develop antigens to generate antibodies specific to species resulting in the development of rapid FTIR spectroscopy-based serotyping (63).

The ability to non-destructively record infrared spectra of intact microorganisms has allowed for the identification of major and minor cellular components. A summary of

various infrared absorption bands assigned to various compounds found in microorganisms (64-67).

Table 2.2. Infrared band assignment of functional groups of biochemical constitutes of microbial cells

Region (cm^{-1})	Biochemical attribute
3500	O-H stretch of hydroxyl group
3200	N–H str (amide A) of proteins
3050-2800	Fatty acid
2955	C–H str (asym) of $–CH_3$ in fatty acids
2930	C–H str (asym) of $>CH_2$
2918	C–H str (asym) of $>CH_2$ in fatty acids
2898	C–H str of C–H in methine groups
2870	C–H str (sym) of $–CH_3$
2850	C–H str (sym) of $>CH_2$ in fatty acids
1800-1500	Amide region
1750-1500	Protein and peptide bonds
1740	>C=O str of esters
1715	>C=O str of carbonic acid
1695, 1685, 1675	Amide I band components resulting from antiparallel pleated sheets and β-turns of proteins
1695, 1637	Amide I (α-helix and β -sheet structures of proteins)
1680–1715	>C=O in nucleic acids
1655	Amide I of α-helical structures
1637	Amide I of β-pleated sheet structures
1550–1520	Amide II
1520	Amid II (peptide bond)
1515	“tyrosine” band
1500-1400	Lipids and proteins
1500-1200	Mixture of fatty acids, proteins and polysaccharide
1468	C–H def of $>CH_2$
1400	C=O str (sym) of COO^-
1310–1240	Amide III band components of proteins
1250	P=O stretching in phospholipids
1250–1220	P=O str (asym) of $>PO2^-$ phosphodiesters
1200–900	C–O, C–C str, C–O–H, C–O–C of carbohydrates, Polysaccharide (O-C and C-O vibrations) – specific to cell membrane
1090–1085	P=O str (sym) of $>PO2^-$
1000-970	Lipopolysaccharides (O-antigen in Salmonella)
900–600	“Fingerprint region”

Different microorganisms will have different microbial cell components and accordingly, these differences are used for FTIR-based phenotypic discrimination between different strains. As the FTIR spectrum reflects the complete biochemical profile of a given

microorganism, it can also be employed for the discrimination between microorganisms of different taxa.

2.5.1.1. Cell wall composition of intact bacteria and yeast

The cell envelop of GN bacteria consists of an outer membrane, peptidoglycan cell wall [2-3 nm comprising of 10% by dry weight of the cell wall (68, 69)] and an inner membrane. The outer membrane is composed of a lipid bilayer composed of phospholipids, glycolipids and LPS. The outer membrane is bound to the peptidoglycan cell wall and is commonly termed the periplasm (densely packed with proteins). All three layers of the cell envelop is typically 4-10 nm thick and encloses the cytoplasm which is approximately composed of 80% water (62, 70) (Table 2.3 and Figure 2.2). Due to the double cell membrane, GN bacteria are inherently more resistant to antibiotics than GP bacteria as their efficacy is dependent on their ability to permeate the cell wall (70-72). In addition, these microorganisms are known to have additional resistance mechanisms such as, production of enzyme that hydrolyze or modifies the antibiotic, efflux pumps or have the ability to lose outer membrane proteins and porins to cause multi-drug resistances (MDR). Changes in the LPS composition can alter the structure of the outer membrane and is responsible for endotoxic shock associated with sepsis (72).

Unlike GN bacteria, GP bacteria do not have an outer membrane, but possess a thick peptidoglycan layer encompassing up to 95% of the cell wall making approximately 40-80% of the dry weight of the cell wall (depending on the species) [Table 2.3; (69)]. The surface features mediate interactions between antibiotics and the environment during an infection. This is especially observed for *S. aureus* where the proteins in the cell envelope drastically change dependent on environmental changes – easily changing in the structure of the outer cell surface. Within the peptidoglycan layer, teichoic acid (absent in GN bacteria) and lipoteichoic acids account for 60% of the cell wall's total mass (62). The cell wall of GP bacteria is approximately 15-80 nm thick depending on the microorganism and growth phase (68) (Table 2.3).

Yeasts are GP fungi with a cell wall that is 20-30% of the total weight of the cell with 85% of the cell wall being carbohydrates and 10-15% proteins. It has a bi-layer structure that

is composed of polysaccharides (alkali-soluble and insoluble β-glucans), mannan and minor chitin (73) with a cell wall thickness of 100-350 nm (74-76).

Table 2.3. Cell wall components of Gram negative and positive bacteria

	Components	Gram negative bacteria	Gram positive bacteria
Cell wall	Cell wall thickness	4-10 nm	15-80 nm
Lipids	Lipid content	15-20%	2-5%
Carbohydrate/Lipids	Lipopolysaccharide content	13%	0%
Carbohydrate/glycerol phosphate	Lipoteichoic acid	Absent	Present
	Number of membranes	2 (7.5-10nm thick outer membrane)	1
Carbohydrates/amino acids	Peptidoglycan content	10-20%	>50%
Protein	Porin proteins	Present	Absent
Protein/Lipid	Protein/lipoprotein content	>50%	0-3%
Carbohydrate/glycerol phosphate	Teichoic acid	Absent	Present

(Table adapted from Hancock, 1998; McGowan Jr, 2006; Poole, 2001; Burattini et al., 2008; Beveridge & Matias, 2006; Davis & Mauer, 2010; Luo et al., 2015; Scorzoni et al., 2017)

Phenotypic diversity of genera and species of yeasts and bacteria are expected to be reflected in differences in their infrared spectra and will be discussed more thoroughly in CHAPTER 4 for relevant species under investigation for the current work.

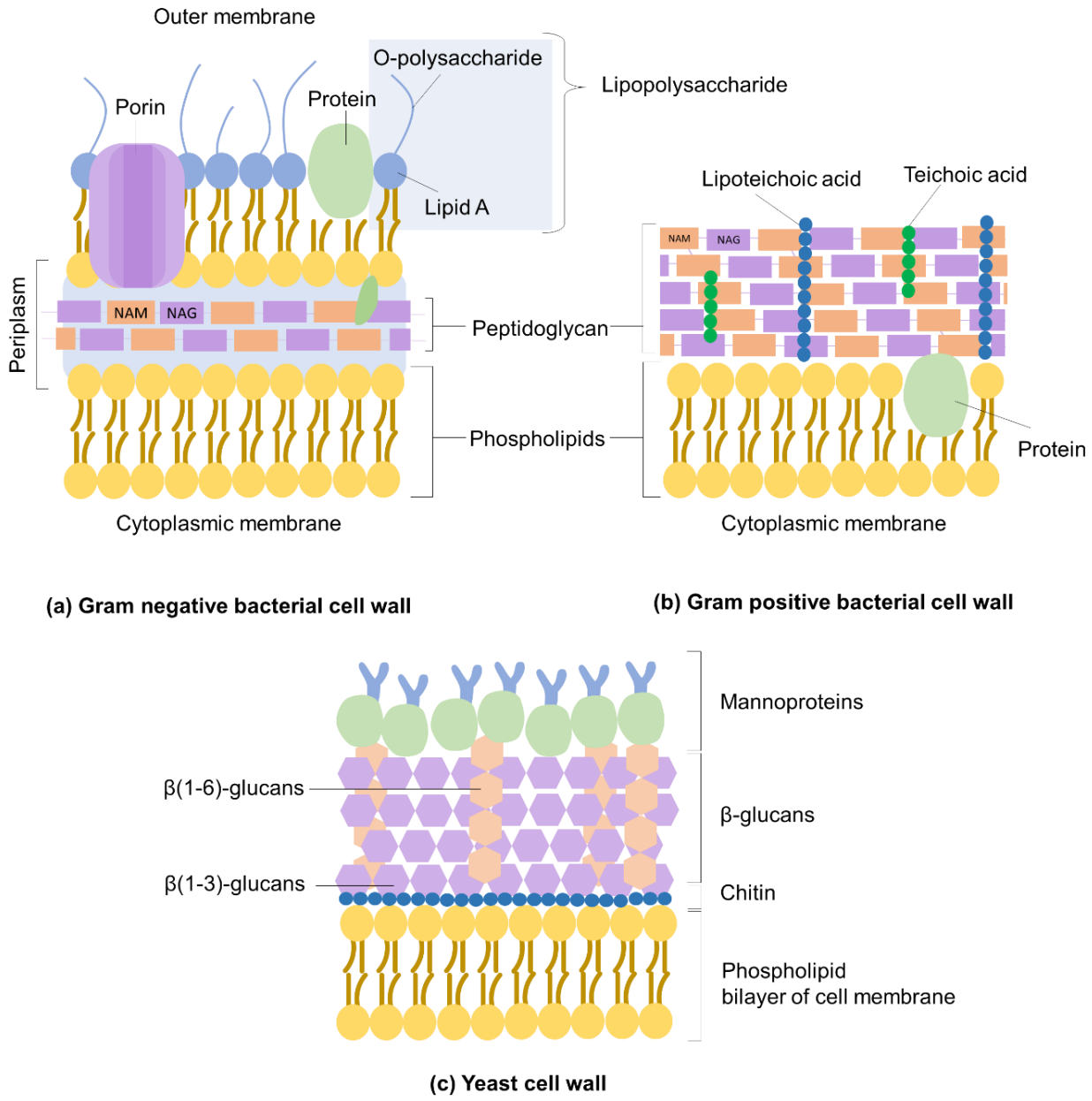


Figure 2.2. Cell wall diagram of (a) Gram-negative bacteria, (b) Gram-positive bacteria and (c) yeast

Accordingly, the cell wall macromolecules contribute to a significant amount of the dry mass of the cell and is the basis for serotyping and species identification based on fatty acid profiles by gas chromatography. While taxonomy undergoes constant changes, infrared spectra of microorganisms recorded under defined growth conditions can be constant. Consistent and repeatable data allows for the development of FTIR spectral libraries for use in microbial identification based on typing through spectral similarities.

Many researchers have successfully investigated the use of FTIR spectroscopy for microorganism identification, classification and strain-typing with the help of chemometric and spectral match (37, 58, 67, 77-86). Although the numerous articles have reported on the applicability of the FTIR spectroscopy-based method for rapid identification of microorganisms, to-date, no study has documented its use in routine microbial identification or have large sample set validation studies.

2.5.2. Significance of changes in biochemical and structural differences of microorganisms as a basis for their identification and discrimination by infrared-based spectroscopic methods

One of the most important criteria for the study of microorganism by FTIR spectroscopy is the acquisition of highly reproducible infrared spectra. For high quality spectral data, spectra acquired consecutively from the same agar plate, from the same strain on multiple days, on different instruments and at different laboratories should be reproducible. This step is crucial as minimal spectral differences from the same strain increase the potential of identifying strains from the same species. High spectral reproducibility also allows the development of shared spectral databases among different laboratories. Spectral reproducibly is dependent on: (i) the signal-to-noise (SNR) of the infrared absorption bands achieved for a given spectrometer, wavenumber (x-axis accuracy), photometric linearity (y-axis), instrument-line-function from different models from the same and different FTIR manufactures, (ii) FTIR spectral profiles of the same strain are affected by culture conditions including media composition, culture time, and incubation environments. Difference in the composition of growth media was found to have the greatest influence on spectral heterogeneity (87, 88). Accordingly, FTIR-based microbial identification methods requires standardization of the growth media composition, maintaining consistent incubation temperature and growth time to achieve highly reproducible metabolic status and (iii) Minimize the effect of sample manipulation and degree of microbial inactivation. Sample inactivation by use of a disinfectant can result in protein denaturation and loss of cell integrity, both of which can alter the spectral profile and potentially remove important components for microbial identification and strain-typing.

Microorganisms are typically grown on agar plates containing defined nutrient compositions for microbial growth. The use of nutrient broth for microbial enrichment is also widely utilized, followed by culturing on nutrient agar plates to confirm microorganism purity. Isolated colonies are acquired from the agar plate for subsequent analysis by most conventional methods and for FTIR-based methods. With regards to the latter, direct analysis of intact microbial cells directly from the agar plate can be carried out. Alternatively, sample treatments can be employed prior to acquiring the FTIR spectrum which include making a microbial suspension, followed by spinning, washing, inactivation of the microorganism and heating subsequent to depositing the sample onto the FTIR sampling surface to form a thin microbial film. Table 2.5 to 2.9 summarizes some of the studies aimed at characterization and differentiation of microorganisms from diverse genera and species by FTIR spectroscopy. It should be noted that sample pre-treatment can have significant effect on the reproducibility of the FTIR spectra.

Once a definitive protocol is established for obtaining highly reproducible spectra of microorganisms, the next step is for the development of robust spectral analysis algorithms to identify spectral features in the infrared spectra. Selective spectral features are characteristic to each strain within a given species that can allow for the differentiation between strains from Gram-stain types, different genera within the Gram type, differences between different species within the same genus and differences between strains within the same species. Accordingly, differences between the strains within the same species allow for the potential utility of FTIR spectroscopy for strain typing. The underlaying changes in the FTIR spectra can reflect a myriad of differences in the strain under investigation. For instance, differences between haemolytic *Bacillus cereus* and non-haemolytic strains can be observed in the FTIR spectra. The question remains in the assignment of the spectral differences between the two groups to biochemical changes reflecting the lack of toxin-producing genes (*ces* and *hbl*) which causes *B. cereus* to become haemolytic (89).

The effect of changing growth media was demonstrated by growing 9 isolates of *Streptococcus* spp. onto two different growth media agar [blood agar (BAP) and tryptone yeast extract agar (TY)] and recording the FTIR spectra of colonies isolated from each

agar growth medium. Examination of the spectra revealed that *Streptococcus* spp. grown on TY, exhibited stronger absorbance bands in the FTIR spectral region between 3000 and 2800 cm^{-1} compared to the spectra of colonies grown on BAP. These bands were tentatively assigned to the presence of amphiphilic molecules from either the cell membrane or cytoplasm (or both) and demonstrates the importance of using a consistent growth medium for microbial identification and strain-typing by FTIR spectroscopy (90, 91). Similar situation was also observed for the use of PFGE for strain typing which initially suffered from lack of standardization and reproducibility, but also became the gold-standard for strain-typing as a result of the development of a highly standardized method for use by microbiology laboratories. Prior to undertaking an extensive evaluation of the capability of FTIR spectroscopy for typing, it is thus essential to develop a standard protocol. Once developed, the true potential of FTIR-based typing method would need to undergo validation studies prior to routine implementation. Furthermore, as FTIR spectral absorptions stem from the biochemical constituents of microorganism (Figure 2.3), fundamental knowledge of species-specific cell composition may aid in the identification of biomarkers that are key to the identification of microorganisms by FTIR spectroscopy.

The extensive overlap of infrared bands of the biochemical components, however, makes it highly challenging to assign infrared bands to specific cellular components serving as biomarkers associated with microbial identification or differentiation. None-the-less, it is possible to identify spectral regions within the FTIR spectra with significant information content that facilitate the discrimination of isolates from family to the serotype level. For example, *Salmonella* serovars (employing O-antigen classification) can be differentiated between each other based on spectral differences in the phospholipid absorption region between 1200 and 900 cm^{-1}, while the use of broader spectral regions was less effective (65). Employing the same spectral region, complete discrimination was also achieved using FTIR spectra of LPS extracts from the same bacteria – demonstrating the importance of their spectral contribution in serotyping by infrared spectroscopy (65). A similar approach was also utilized in serotyping *K. oxytoca* (92). Table 2.4 summarizes the results of additional examples of serotype differentiation by FTIR spectroscopy.

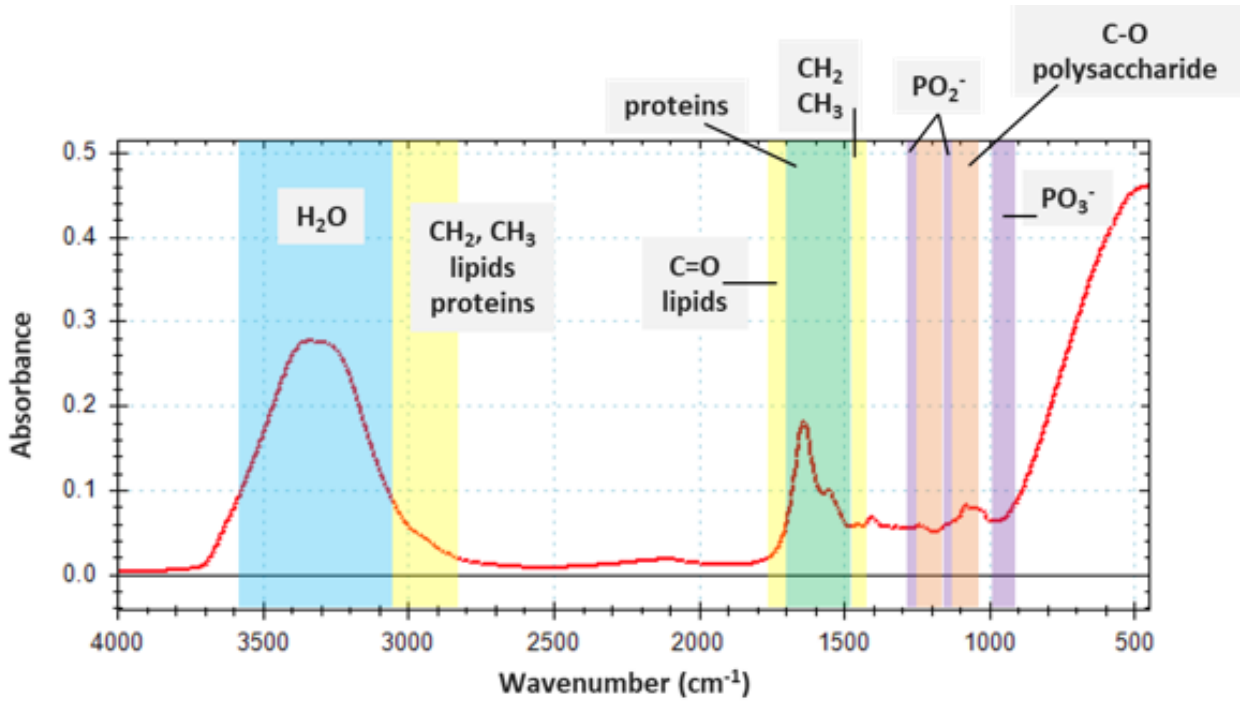


Figure 2.3. Typical FTIR spectra of a wet microorganism and biochemical components associated with the mid-infrared wavenumbers

Table 2.4. Summary of studies for FTIR spectroscopy in microbiology for species identification and strain typing by macromolecule extraction

Description	Sample size	Biochemical key components	Data processing	Regions	Data analysis	Results	Ref.
Staphylococcus aureus serotyping by capsule extraction Capsular polysaccharides are important virulence factors	Capsule serotypes: CP5 (23) CP8 (27) NT (37) Total: 87	Polysaccharides (1200-900 cm^{-1}: C-O-C and C-O-P stretching)	2nd derivative (Savitzky-Golay) Vector normalization	Feature selection (COVAR algorithm) of spectral windows (3000-2800 cm^{-1}, 1800-500 cm^{-1}) – actual regions not disclosed	HCA ANN with classification by WTA (winner takes all)	98.2% correct serotype identification by ANN	(93)
Listeria monocytogenes serotype identification (12 serovars)	Database (106) External validation (166)	O-antigen (polysaccharide of the LPS structure), teichoic acids, H-antigen	1st Derivative Savitzky-Golay	1200-900 cm^{-1} and 1800-1400 cm^{-1}	HCA ANN	98.8% correct for O-antigen (serogroup) 91.6% H-antigen (serovar) and 40 of 41 epidemic serovar 4b were identified (outbreak)	(63)
Salmonella enterica serotyping by LPS extraction (6 serotypes)	6	Lipopolysaccharides (also known as glycolipids and LPS)	No preprocessing techniques used	1200-900 cm^{-1}	CVA PCA	100% classification with LPS extract and 47% for intact cells	(65)

Abbreviations: HCA: hierarchical cluster analysis; ANN: artificial neural network; CVA: canonical variate analysis; PCA: principal component analysis

2.5.3. Sample preparation of microorganisms for FTIR spectral acquisition

Traditionally, colonies are removed from an agar culture plate (after a 16-24 h incubation period at a defined temperature) transferred into a vial, rinsed, and inactivated prior to smearing on an infrared (IR) transmitting substrate (ZnSe or Si), left to dry at room temperature or heated to expedite drying. More recently, the use of attenuated total reflectance (ATR)-FTIR spectroscopy and specular reflectance FTIR spectroscopy for acquiring spectra of microorganisms has been reported. Figure 2.4 shows three different modes of acquiring spectra from different FTIR sampling accessories; for transmission measurements, ZnSe or Si are used as an IR optical window, specular reflectance uses a silver or gold-coated glass slide and, by ATR using a diamond or Si ATR internal reflection crystal. Not as common, diffuse reflectance mode of spectral acquisition have been studied (Table 2.5), samples are freeze-dried and mixed with KBr (transparent to IR light) and placed in a diffuse reflectance accessory where the IR energy undergo absorption-reflection. Distortion of the spectral bands can result from the use of this method to acquire IR spectra of microorganisms to the sample and measures the absorption energy to reflect it back over a large angle (69).The use of each of these sampling methods for recording FTIR spectra of microorganisms are summarized in Table 2.5 to Table 2.9. The diversity of sampling methods coupled with the use of different growth media and growth parameters makes it difficult to evaluate the performance of an FTIR-based microbial analysis approach.

2.5.4. Modes of FTIR spectral acquisition for microorganisms

Traditionally IR spectra are recorded between 4000 and 650 cm^{-1}, the spectral region with substantive information related to the differences in biochemical composition and water content are 3800-2800 and 1800-650 cm^{-1}. IR spectra recorded using FTIR spectroscopy in the transmission mode (most commonly used), require ~10-60 µg of cells (based on wet weight) (69) spread uniformly as a thin-dried film over an area of ~3x3 mm onto an IR-transparent window where the IR beam passes through the sample to reach the detector to produce a spectrum. Although many studies have successfully utilized this method as it offers a high SNR, there are reported variation between spectra due to

variability in sample thickness, uniformity, and sample pre-treatment procedures used to prepare the microbial films (Table 2.5).

Specular reflectance (Figure 2.4) has a similar sample preparation protocol used for recording transmission-based FTIR spectra. The sample is placed on an IR reflective substrate (or window) and the IR beam effectively passes through the sample, is reflected back prior to recording the FTIR spectrum. Accordingly, sample thickness must be ½ of that employed for transmission measurements.

Due to ATR-FTIR spectroscopy's independence of sample thickness (after coverage of the crystal with ~1-2 μm-thick sample), it is the most versatile method of spectral acquisition of intact cells of microbes. Most ATR-FTIR-based studies deposit the microbial colonies directly onto the ATR crystal (or element), wait for the water in the sample to evaporate (~10-30 minutes) to form a dry film on the crystal, then record the ATR-FTIR spectrum with very high SNR. ATR-FTIR spectra of microbial cells can also be recorded without the need of drying, however, sample dilution by water in the colony was observed.

2.5.5. FTIR spectral processing

Spectral processing of highly reproducible FTIR spectra of microorganisms facilitates identification of spectral features responsible for discriminating among different strains. Various processing techniques can be carried out such as, baseline correction to compensate for baseline drift between a set of spectra, spectral smoothing and transformations. The most common processing technique is by calculating the 1^{st} or 2^{nd} derivatives of spectral sets. Derivative spectra yield a mathematically based resolution enhancement that helps delineate spectral features of overlapping bands. It should be noted that derivative spectra can produce some increase in spectral noise and should be applied when the absorbance spectra (raw spectra) have high SNR. IR spectra recorded from microbial films using transmission or specular reflectance spectroscopy can have widely different absorbances due to variability in film thicknesses.

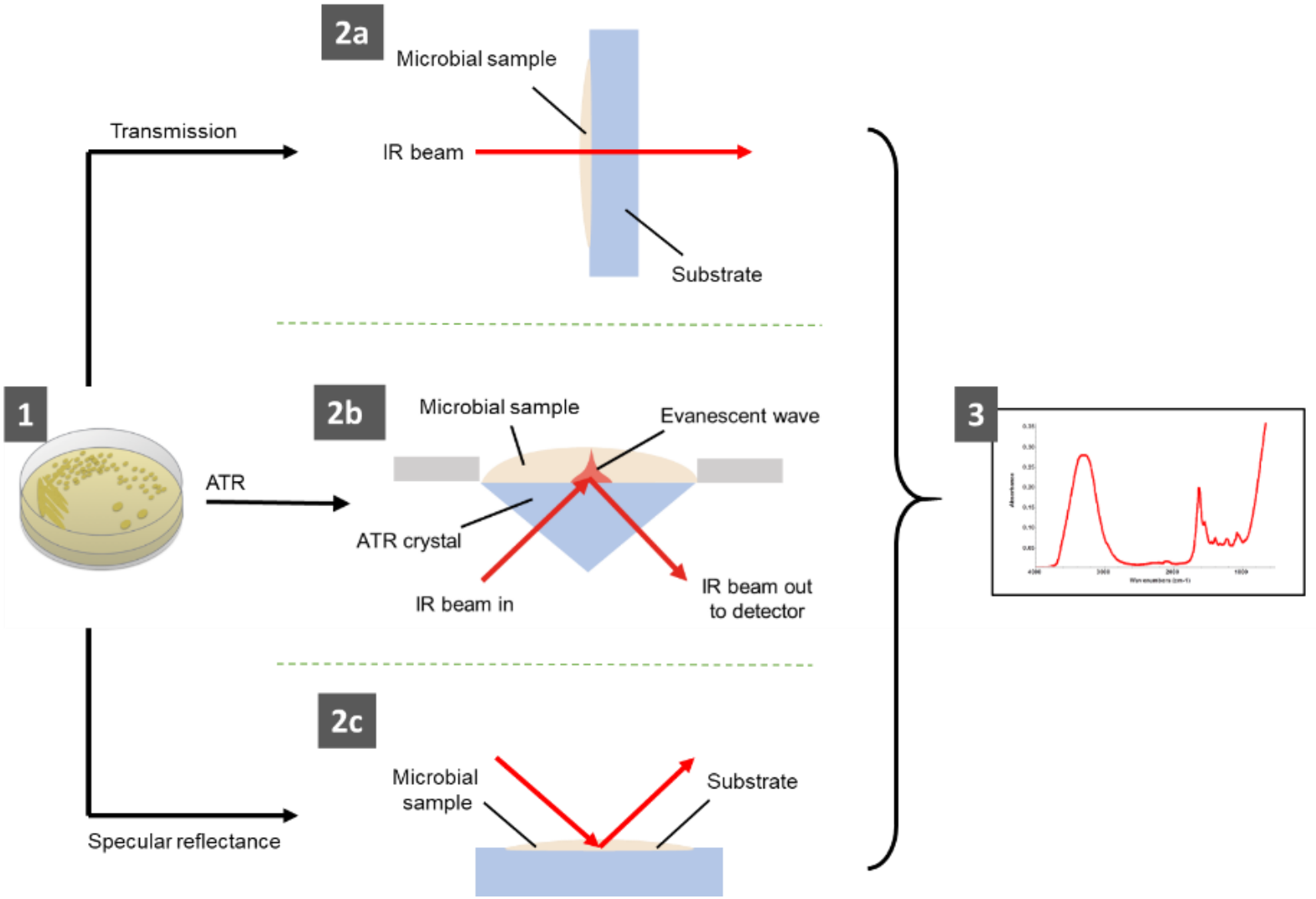


Figure 2.4. Three different modes of FTIR spectral acquisition illustrated for spectral acquisition of microorganisms from culture agar media. Spectral acquisition modes: (2a) Transmission, (2b) attenuated total reflectance (ATR), and (2c) specular reflectance

This can be compensated by employing a spectral normalization algorithm. Film thickness heterogeneity and dryness can also result in significant light scattering that can produce optical artifacts. Several algorithms are also routinely employed to compensate for this (e.g., scattering correction).

In the case of ATR-FTIR spectra of microbial colonies directly placed on the ATR sampling surface, 1st derivative calculation is employed for baseline correction and resolution enhancement of the absorbance bands followed by vector normalization to compensate for the variability in colony cell density among the samples (69).

2.5.6. Data mining and analysis (chemometrics) of FTIR spectra for identification and strain typing of microorganisms

2.5.6.1. Spectral library search approach

Spectral libraries are created comprising of spectra of pure and well characterized microorganisms and serve as a reference database by which a spectrum of an “unknown” microorganism can be compared against. Several algorithms can be utilized individually or in tandem to measure the degree/extent of spectral similarity of the spectrum of the “unknown” microorganism to one or more of the spectra in the reference database. Spectral similarity can be expressed in terms of cosine, Pearson or Euclidean distances. Distance is defined by the dissimilarity of the spectra and is represented as a dimensionless measurement (or the area between two overlay spectra of the spectral library/reference database and spectra of unknown isolate). Each distance equation computes a number that is interpreted differently, for example, Pearson’s product distance is a value between 0 and 2000 where 0 is 100% match to a spectrum in the spectral library and cosine ranges from -1 to 1 where 1 is 100% match (87, 94). Detailed mathematical explanation, examples and references therein can be found in (95). It is important to note that spectral distance can also help in the identification of spectral artifacts among replicate infrared spectra of the same strain (96).

In an early study conducted by D. Helm et al. (94), using the Pearson’s product moment covariance distance match search, they concluded that increased spectral representation of each species in the spectral library is required for improved identification rates. The

researchers also considered that while populating the spectral library, more spectral overlap will take place among different species, making it difficult to identify an unknown based on its IR spectrum. This drawback can be addressed by calculating mean spectra for each species revealing subtle features between different species (60). Although some success was achieved by this approach (61), many of these studies have limited number of spectra from multiple species in the spectral libraries (66). Furthermore, these approaches have not been subjected to extensive internal or external validation studies.

2.5.6.2. Principal component analysis (PCA) (unsupervised)

PCA is a visualization method of displaying large and complex data contained in a matrix (e.g., spectral groups representing different microbial species) where natural clustering is observed in a 2-dimentional or 3-dimentional space. PCA evaluates the differences between the data in addition to identifying which variables contribute to the major differences between the data and whether there is a correlation between the groups or not. The principal component (PC) scores explain the variabilities in the data. This method of analysis is commonly used in the analysis of FTIR spectra of microorganisms. H. Haag et al. (87) demonstrated that the use of PC scores that accounts for 90% of variances in the spectral data yielded discrimination of species and 89% correct identification of unknowns. Other studies have also used selected spectral regions as an input for PCA to identify microorganisms (97). It is possible that subtle variability between FTIR spectra of certain microorganisms may not contain all the information needed to discriminate between them. Additional approaches to data mining may be required and this will be further investigated in the thesis.

2.5.6.3. Hierarchical cluster analysis (unsupervised)

HCA is an unsupervised clustering method to determine the relationship between spectral groups and is visualized by plotting the similarity distance in the form of a dendrogram and distance (95).

For example, using HCA, an FTIR spectral data set of 332 yeasts (food origin) belonging to 12 genera was separated into 22 major clusters. However, HCA was not effective to completely separate the 12 genera from one another (66). This example demonstrates

the need for alternative spectral analysis approaches to determine if additional spectral information can be found to discriminate between all 12 genera.

Additional studies have utilized FTIR spectroscopy and HCA to correctly strain-type different microorganisms (97-99). A high discriminatory power of 0.983 [based on Simpson's index of diversity (SID)] for unrelated *S. aureus* strains was achieved based on differences in their FTIR spectral profiles where *spa* typing and PGFE (gold standard) obtained roughly 0.976 SID discriminatory power (99). This study encompassed 70 isolates and demonstrated the applicability of FTIR spectroscopy as a potential tool for outbreak investigation. Unlike genotypic techniques such as PFGE, FTIR spectroscopy discrimination among isolates is based on changes in the biochemical constituent on surface glycopolymers, capsular serotypes and major cell-wall constituents (65, 99). Moreover, differences in the FTIR spectra of a microorganism can be indicative of changes in the metabolomic status of each isolate.

2.5.6.4. Discriminant Analysis (supervised)

Chemometrics is a word primarily used for spectral data to qualify and/or quantify chemical data mathematically (linear algebra). Data are typically used for exploratory analysis, regression, and classification. The following sections will discuss various chemometric techniques for identification and strain-typing of microorganisms from FTIR spectra.

Partial least squares (PLS) discriminant analysis (DA) is a supervised method performed in junction with PCA (unsupervised) to enhance the discrimination between natural clusters observed by rotating between the PC. This method of DA is commonly used to group the spectra of unknowns with defined classes and assigning it to one of the classes. A drawback of this method stems from the possibility of assigning a spectrum of an unknown microorganism (not present in the database) to one of the defined classes. The "outlier" samples, defined as samples whose identity may not be accurate or is not represented in the model, will be classified into one of the defined classes. Interestingly, various studies have successfully applied PLS-DA for identifying microorganisms and claimed to have higher classification power than other models such as Soft Independent Modeling of Class Analogy (SIMCA) (82, 100).

By comparison, using SIMCA algorithms coupled with PCA differs from PLS-DA by either classifying the unknowns into predefined classes achieved by high probability or as an outlier (low probability). SIMCA takes the average residual variance of the data of a defined class to the residual variance of the data from the spectrum of the unknown (goodness-to-fit of the PCA model). If the spectrum of unknown is not represented in one of the defined classes, it will be predicted as an outlier or class-not-defined. In addition, SIMCA allows for data quality assessment as the discriminatory power of this approach is highly sensitive to the raw data (microorganism spectra). The approach is also not limited to the number of defined classes, which makes it desirable for microorganism classification using spectral data of many species and does not contribute to collinearity and chance classification.

Various studies have applied SIMCA for identifying microorganisms and strain-typing (64). In a study, 100% classification of *Salmonella* serovars (Enteritidis, Typhimurium, Heidelberg, Kentucky and Anatum) was achieved using the SIMCA approach by employing a narrow spectral region between 1000 and 970cm^{-1} (tentatively assigned to variability in LPS composition). The study explored the effects of using different growth media; plate count agar, xylose lysine deoxycholate agar (XLD – inhibits GN bacteria and selects for *Salmonella*) and Miller-Mallinson agar (MMA – selective for non-salmonellae lactose fermenters). Although successful serotype classification of *Salmonella* serovars was achieved; selective media such as MMA performed the poorest with plate count agar and XLD performing similarly. This study further proves the effects of culture media on the microorganism's biochemical composition and its distinctive effects on the FTIR spectra of the microorganisms (64). It should be noted that the use of SIMCA requires many classes (>10) where R. G. Saraiva et al. (101) attempted the use of SIMCA for discrimination between 3 different species, however this approach failed to get adequate results.

Other analysis method such as k-nearest neighbors, artificial neural networks (ANNs), support vector machines (SVM) and canonical variate analysis (CVA) are also widely used for spectral analysis. The methods are well described elsewhere and all references therein (Ghetler, 2010).

2.5.7. Summary of current studies in literature for FTIR spectroscopy-based microorganism identification, strain-typing and differentiation

L. Mariey et al. (83) have summarized many of the studies that have employed FTIR spectroscopy from before 2001, almost 2 decades have passed since the literature review where more recent studies are presented in Table 2.5 and were discussed throughout the current review. Most studies (*n*=26) have employed FTIR spectroscopy in the transmission mode of spectral acquisition with 5 studies using ATR, 4 using diffuse reflectance, 1 for specular reflectance and 2 for microscopy. In all 33 published FTIR spectroscopy work presented in Table 2.5 through Table 2.9, only 8 were for classification and identification of microorganisms, while the remaining published work employed FTIR spectroscopy for discriminating among a limited selection of microorganism at the species level.

Table 2.5. (Part 1 of 5 tables) Summary of current studies (past 2 decades) in literature for FTIR spectroscopy for microbial identification, strain-typing and differentiation

Research details (no. species)	No. of isolates Reference/Unknown	Sample cultivation	Spectral processing	Data analysis method	Results	Ref.
Species identification						
ATR mode of spectral acquisition						
S. aureus, CoNS, Non-*Staphylococcus*	35/58 (*Staphylococcus* only)	Culture plate, wash, spin, wash, spin, vacuum dried with desiccant	Vector normalized, Savitzky-Golay, 1st Derivative	HCA, PLSR, SIMCA	65.5% correct species by HCA, 98.3% correct species by PLSR	(102)
S. aureus	1/N/A	Direct from spiked blood	Baseline corrected, 1st derivative (Savitzky-Golay)	SPA-LDA, GA-LDA	Successful exploratory study	(79)
Cryptococcus species (2 species)	22/6	Culture plate, covered with aluminum foil	Baseline correction	PCA, LDA, QDA, SPA-LDA, SPA-QDA, GA-LDA, GA-QDA	Species discrimination and classification was achieved, best results with GA-QDA	(103)
Candida species (5)	NA/82	Direct culture plate	Standard normal variate, 3rd derivative, Savitzky-golay (37-point) smoothing	PCA, PLSDA	99.6%	(104)
Listeria species (2 species) and *Salmonella enterica* serovars (3 serovars) (Food)	14/NA	Culture plate, airdried (10 seconds)	Standard normal deviate	PCA	Successful discrimination between species and serovars	(105)
Yeasts (12 genera and 65 species)	263/318	Direct from culture plate	1st Derivative, Vector normalize	HCA, PCA, Forward search	99.7%	(81)

Table 2.6. (Part 2 of 5 tables) Summary of current studies (past 2 decades) in literature for FTIR spectroscopy for microbial identification, strain-typing and differentiation

Research details (no. species)	No. of isolates Reference/Unknown	Sample cultivation	Spectral processing	Data analysis method	Results	Ref.
Species identification						
Transmission mode of spectral acquisition						
Coryneformbacteria (46 genera and 220 species)	730/544	Culture plate, distill water suspension, 42ºC heat dried (1 h)	1st Derivative	Spectral library search	95.4% genus, 87.3% species	(106)
Listeria (5 species)	243/277	Culture plate, distill water suspension, 42ºC heat dried (1 h)	Savitzky-Golay, 1st Derivative	HCA, ANN, Univariate	96.3% (ANN), 88.9% (Univariate)	(107)
Listeria (5 species)	25/NA	Culture plate, distill water suspension, 42ºC heat dried (1 h)	1st Derivative	Spectral library approach	92.8%	(108)
BCC	185/10	Plate cultures, suspension, 55ºC heat dried (45mins)	SNV, Savitzky-Golay, 1st derivative	PCA, SIMCA, PLSDA, ANN	90.0%	(109)
Lactic acid bacteria (9 genera and 92 species)	379/85	Plate cultures, suspension, 40ºC heat dried (45mins)	1st Derivative, Vector normalize	ANN	93.2% species, 97.1% strain	(110)
Bacteria (93 species in 4 taxonomic groups)	In-house database (3502 spectra)/40	Suspension, 42ºC heat dried (45 mins)	1st derivative (*Staphylococcus*) and 2nd derivative for other	Spectral library approach	75.0%	(91)

Table 2.7. (Part 3 of 5 tables) Summary of current studies (past 2 decades) in literature for FTIR spectroscopy for microbial identification, strain-typing and differentiation

Research details (no. species)	No. of isolates Reference/Unknown	Sample cultivation	Spectral processing	Data analysis method	Results	Ref.
Species identification						
Transmission mode of spectral acquisition						
Filamentous fungi (43 genera and 140 species)	288/105	Extraction of mycelial by washing and centrifugation, resuspension, dried under mild vacuum (1 h)	Baseline corrected (64 points), 2nd derivative (9-point Savitzky-Golay), vector normalized	PLS-DA	99.17% genus, 92.3% species	(82)
Bacteria: GP (3 genera, 3 species), GN (3 genera, 4 species); Yeasts: (*Candida* – 6 species)	121/121	Positive blood culture, dilute and cultured on agar (6-8h), stamped and airdried (15mins)	1st and 2nd derivatives	HCA, LDA, ANN	98.3% species	(67)
Transmission-microspectroscopy mode of spectral acquisition						
Candida species (3 species)	30/NA	Culture plate, Dried by desiccant and vacuum	Elastic correction, 2nd Derivative, Vector normalized	HCA	Successful discrimination between species	(96)
Diffuse reflectance mode of spectral acquisition						
E. coli, *S. aureus*, *C. albicans*	NA/3	Culture plate, dried and homogenized with KBr	Baseline corrections, normalized, smoothed, 2nd derivative	HCA	Successful discrimination between species	(111)
Specular reflectance mode of spectral acquisition						
Candida species (3 species)	169/45	Culture plate, suspension, dry under hot air 60ºC for 3 mins.	Baseline corrected, Normalized, 1st Derivative	PA, SIMCA	93.4%	(112)

Table 2.8. (Part 4 of 5 tables) Summary of current studies (past 2 decades) in literature for FTIR spectroscopy for microbial identification, strain-typing and differentiation

Research details (no. species)	No. of isolates Reference/Unknown	Sample cultivation	Spectral processing	Data analysis method	Results	Ref.
Strain typing						
ATR mode of spectral acquisition						
Candida species (5 species)	5/20	Direct from culture plate	Normalized	PCA	Strain-typing of samples to patients were not consistent for some patients going through antifungal treatment, some patients displayed similar clustering as PFGE	(113)
Salmonella serovar (6 serovars)	6/8	Culture plate, suspension, dried with 50% EtOH under vacuum (25ºC for 3-5 mins)	2nd Derivative Savitzky-Golay, Normalized	SIMCA, PCA	100%	(114)
Single porin *E. coli*	8/NA	Direct from culture plate and air dried for 20-30 mins	Standard normal variate, 2nd Derivative Savitzky-Golay	PCA, HCA	Successful intra-species discrimination of porin expression profile	(101)
Specular reflectance mode of spectral acquisition						
E. coli O157:H7	30/30	Culture plate, wash, suspend, wash, suspend and placed onto gold slide	Baseline corrected, Smoothed, 1st/2nd derivative (9-point Savitzky-Golay), Normalized	HCA, CVA	97%	(115)

Table 2.9. (Part 5 of 5 tables) Summary of current studies (past 2 decades) in literature for FTIR spectroscopy for microbial identification, strain-typing and differentiation

Research details (no. species)	No. of isolates Reference/Unknown	Sample cultivation	Spectral processing	Data analysis method	Results	Ref.
Strain typing						
Transmission mode of spectral acquisition						
Patient cluster of yeast isolates	79/NA	Dried by desiccant and vacuum	2nd derivative (9-point Savitzky-Golay)	HCA, Feature selection	9 patients, 1 of 79 isolates misclassified with a different patient	(116)
Candida species (C. *albicans*, C. *glabrata*, C. *parapsilosis*)	40/NA	Culture plate, Dried by desiccant and vacuum	Elastic correction, 2nd Derivative, Vector normalized	HCA	Successfully discriminated between 12 patients based on the 3 species	(96)
S. aureus	NA/70	Culture plate, distill water suspension, 40ºC heat dried (40 mins)	Vector normalized, Pearson's product moment correlation coefficient	HCA, PCA, ANN, LDA, SVM	Comparable discriminatory power for gold standard PFGE	(99)
Klebsiella spp.	NA/10	Plate cultures, suspension, dried via desiccator over a drying agent	2nd derivative (9-point Savitzky-Golay), Vector normalized	HCA	Successful discrimination between strains	(98)
Klebsiella spp.	68/NA	Culture plate, EtOH suspension, water suspension, dried (37ºC for 20 mins)	Vector normalized	HCA	92.6% agreeance with WGS clusters	(92)
P. aeruginosa, K. pneumoniae, E. cloacae., A. baumannii	156/NA	Culture plate, EtOH suspension, water suspension, dried (35ºC for 25 mins)	Vector normalized, 2nd Derivative	HCA	Successful discrimination between species and those belonging to the same clones	(117)

2.6. Future use of ATR-FTIR spectroscopy in microbiology

MALDI-TOF MS has changed the clinical microbial identification landscape leading to future acceptance of spectroscopic methods as alternatives to conventional identification methods such as carbon assimilation tests. With a simpler sample preparation compared to MALDI-TOF MS, FTIR spectroscopy is poised for expanding the current array of studies including identification of microorganisms from biofluids, positive blood culture and mixed colonies (65).

2.6.1. Identification of microorganisms isolated from positive blood cultures

Rapid microbial identification directly from blood is highly sought out and is the subject of extensive research. Kits for rapid recovery of microorganisms from positive blood cultures for identification by MALDI-TOF MS have been developed recently. Some success has been achieved for the isolation of microorganisms from positive blood cultures, urine and other bodily fluids analysis by MALDI-TOF MS. The MBT Sepsityper ® kit (Bruker Daltonik GmbH, Bremen, Germany) requires trained laboratory technicians to extract 1mL of fluid from a positive blood culture tube and transferring the aliquot to a sterile 2ml tube and adding 200µL of lysing buffer. The sample is then vortexed for 30 seconds and centrifuged for 1 minute at 13000 rpm; the supernatant is then removed followed by a washing step. The remaining pellet is washed with 1mL of washing buffer and centrifuged for 1 minute at 13000 rpm followed by the removal of the supernatant (118). Based on published literature, the MBT Sepsityper ® kit (Bruker Daltonik GmbH, Bremen, Germany) has reported low accuracy readings for anaerobic bacteria and yeasts (119), additional washes may be needed to obtain reliable species identification for yeasts (Table 2.10).

MALDI-TOF MS (Bruker Daltonik GmbH, Bremen, Germany) identification scores (reflecting the accuracy of microorganism identification) are highly dependent on the concentration of bacteria and yeasts being spotted on the target plate. Less bacteria present was observed to have lower scores of <1.7 on the Bruker system, meaning a smaller number of mass peaks produced by the sample, making the result less reliable.

From the observations of T. J. Gray et al. (120), a minimum of 10^6 CFU is needed to obtain a reliable spectral score especially for the *Streptococcus viridans* group (121-123).

Table 2.10. MALDI-TOF MS-based species identification performance from positive blood culture using in-house methods and the MBT Sepsityper ® kit (Bruker Daltonik GmbH, Bremen, Germany) kit

			Results	
Reference	Protocol starting from positive blood culture	Sample size	Genus	Species
(Gray, Thomas, Olma, Iredell, & Chen, 2013)	5ml placed into a blood serum separator (15mins), decant-wash-centrifuge (5mins), spot, dry	292/318 (poly colonies)	(*Enterobacteriaceae*) NFGNB (93.3) NFGNB (84.0) Other GN (75.0) Overall: 91.8	(*Enterobacteriaceae*) NFGNB 90.9% NFGNB (80.0%) Other GN (75%) Overall: 89.4
(Schieffer et al., 2014)	SepsiTyper kit	325/411 (No ID - poor protein extraction - mixed organisms)		*Enterobacteriaceae* 98.9 Other GN GN (98.9) GP: 82.9 YT: 50.0 AN: 0.00 Overall: 92.6
(Stevenson, Drake, & Murray, 2010)	Laboratory developed method: Serum separator, centrifuge, wash to remove red blood cells, resuspend, wash, lysed, incubated for 10 mins, washed, resuspend, wash, ethanol	138/212		Overall: 96.2
(La Scola & Raoult, 2009)	A: spin, decant, resuspend, spin, add acetonitrile, trifluoroacetic, incubate, spine, spot B: spine, decant, resuspend, spin, add formic acid to pellet, incubate, add acetonitrile, vortex, spin, decant, spot	599/621 (mixed culture)		A: GN: 94.0 GP: 37.0 Overall: 59% B: GN: 87 GP: 67 Overall: 76%
(Verroken et al., 2015)	Positive blood culture, culture for 5h on Columbia blood agar or brucella (AN), spot target plate as usual	896/913	GN: 25 GP:61	GP:82.2 GN:90.4 *Enterobacteriacea* 92.7 Non-fermenters: 94.1 AN: 20 YT: 0 Overall: 81.1
(Juiz et al., 2012)	Sepsityper	85	GN: 96 GP:96.7 Overall:96.5	GN:96 GP:86.9 Overall:85.2
(Buchan, Riebe, & Ledeboer, 2012)	Sepsityper	146	GN: 97.6 GP: 98.1 YT: 0 (polymicro)	GN:95.1 GP:93.3 YT: 0 (polymicro)

Moreover, MALDI-TOF MS requires pure colonies and discrepant results were reported in various studies and attributed to inadequate sample rinsing. Those studies recommended carrying out a Gram stain prior to MALDI-TO MS analysis from direct blood cultures – adding an additional step to an already lengthy protocol (119, 121, 124).

Additional studies revealed that 3 and 4 h incubation periods of blood cultures had poor results for the identification of species belonging to GP bacteria and sub-par results for GN bacteria (123, 125). According to A. Verroken et al. (123), with an addition step of culturing positive blood cultures onto growth medium for 9 h, resulted in sufficient identification for GP bacteria while 2 h incubation time is acceptable for GN bacteria. Another drawback of microbial identification obtained from positive-blood culture is the possible presence of polymicrobial strains (126).

Overall, direct isolation of microorganisms from positive blood cultures still requires at least 30 minutes of washing and is labor intensive, making it impractical for clinical routine workflow analysis (unless it is automated). As stated above, it may also require an additional culturing and incubation step (2 and 9 hours) from blood culture to yield enough biomass for MALDI-TOF MS analysis.

Our laboratory was the first to demonstrate the utility of FTIR spectroscopy as a method of identification for bacteria isolated from positive blood cultures (unpublished work). This work is in its early stages of development and will require standardization. An FTIR-based method should greatly reduce time, consumables and costs for large and small laboratories, and could be easily integrated into the laboratory workflow. Microcolonies of bacteria isolated after incubation (for ~4 hours on agar culture medium) of microorganisms isolated from positive blood cultures could shorten the analysis time and have sufficient microbial mass for FTIR analysis (11, 67). The issue with microcolonies is that not all microorganisms grow at the same rate, yeasts for example are slow growers (11, 67). Studies have shown that bacteria grown after 8 h of incubation on agar culture media yielded 98.3% correct species identification for 121 samples of common bacteria and yeast isolates by FTIR microscopy (59). It is also the only reported available study for the positive blood culture and subculture onto agar for a short period of time to obtain microcolonies. L.-P. Choo-Smith et al. (59) conducted a study using the linear discriminant analysis and ANNs with leave-one-out validation and achieved 98.3% (119/121) correct species identification of 14 species and discriminating out the CoNS group (a total of 6 genera encompassing both bacteria and yeasts of clinical relevance) (59).

Other studies have reported substantial variability in the FTIR spectra between microcolonies of species within the genus *Listeria* – limiting the identification of the microorganism (108). Intra-species variability has been attributed to variability of cell growth phase of each cell (98, 127) and D. Toubas et al. (96) demonstrated that for the conventional culturing method, the spectral distances between replicate spectra of *Listeria* species were 0.15, whereas, using spectra from microcolonies of the same samples (with reduced culturing time) resulted in an increase in spectral and colony heterogeneity of 0.45 and 0.85. FTIR-based microbial identification by microcolonies may be feasible for routine identification but may be limited to specific species.

Although studies on microcolonies have been discussed, no study have truly evaluated or standardized this approach for rapid microbial identification by FTIR spectroscopy. It is possible that the heterogeneity of the replicate spectra can be reduced by increasing the spectral databases and employing different data-mining techniques such as ANNs.

2.7. GENERAL SUMMARY OF LITERATURE REVIEW

The purpose of the current review is to demonstrate the multidisciplinary research approach which encompasses the use of vibrational spectroscopy, microbiology, microbial epidemiology, molecular biology, biochemistry, genetics, multivariate statistics, chemometrics and microbial isolation techniques to establish a reliable ATR-FTIR spectroscopy-based workflow for use by clinical and food microbiology laboratories. In this multi-faceted field of study, spectral acquisition requires microbial culture and spectral processing standardization to ensure inter and intra-laboratory reproducibility.

2.8. REFERENCES

1. Porter J. 1976. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. Bacteriological reviews 40:260.
2. Abel K, Deschmertzing H, Peterson J. 1963. Classification of microorganisms by analysis of chemical composition I.: Feasibility of Utilizing Gas Chromatography. Journal of Bacteriology 85:1039-1044.
3. Nakase T, Komagata K. 1971. Significance of DNA base composition in the classification of yeast genus Torulopsis. The Journal of General and Applied Microbiology 17:161-166.
4. Norris J. 1964. The classification of Bacillus thuringiensis. Journal of Applied Bacteriology 27:439-447.
5. Rhodes ME. 1965. Flagellation as a criterion for the classification of bacteria. Bacteriological reviews 29:442.
6. Whittaker RH, Margulis L. 1978. Protist classification and the kingdoms of organisms. Biosystems 10:3-18.
7. Abbott SL. 2011. Klebsiella, enterobacter, citrobacter, serratia, plesiomonas, and Other Enterobacteriaceae, p 639-657, Manual of Clinical Microbiology, 10th Edition. American Society of Microbiology.
8. Iversen C, Mullane N, McCardell B, Tall BD, Lehner A, Fanning S, Stephan R, Joosten H. 2008. Cronobacter gen. nov., a new genus to accommodate the biogroups of Enterobacter sakazakii, and proposal of Cronobacter sakazakii gen. nov., comb. nov., Cronobacter malonaticus sp. nov., Cronobacter turicensis sp. nov., Cronobacter muytjensii sp. nov., Cronobacter dublinensis sp. nov., Cronobacter genomospecies 1, and of three subspecies, Cronobacter dublinensis subsp. dublinensis subsp. nov., Cronobacter dublinensis subsp. lausannensis subsp. nov. and Cronobacter dublinensis subsp. lactaridi subsp. nov. International journal of systematic and evolutionary microbiology 58:1442-1447.

9. Teixeira LM, Merquior VLC. 2013. Enterococcus, p 17-26, Molecular Typing in Bacterial Infections. Springer.

10. Al-Mutairi MF. 2011. The incidence of Enterobacteriaceae causing food poisoning in some meat products. Advance Journal of Food Science and Technology 3:116-121.

11. Maquelin K, Kirschner C, Choo-Smith L-P, Ngo-Thi N, Van Vreeswijk T, Stämmler M, Endtz H, Bruining H, Naumann D, Puppels G. 2003. Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures. Journal of clinical microbiology 41:324-329.

12. Becker K, Skov RL, von Eiff C. 2015. Staphylococcus, Micrococcus, and other catalase-positive cocci, p 354-382, Manual of Clinical Microbiology, Eleventh Edition. American Society of Microbiology.

13. Moran GP, Sanglard D, Donnelly SM, Shanley DB, Sullivan DJ, Coleman DC. 1998. Identification and Expression of Multidrug Transporters Responsible for Fluconazole Resistance in Candida dubliniensis. Antimicrobial Agents and Chemotherapy 42:1819-1830.

14. Howell SA, Hazen KC. 2011. Candida, Cryptococcus, and other yeasts of medical importance, p 1793-1821, Manual of Clinical Microbiology, 10th Edition. American Society of Microbiology.

15. Hale TL, Sansonetti P, Schad P, Austin S, Formal S. 1983. Characterization of virulence plasmids and plasmid-associated outer membrane proteins in Shigella flexneri, Shigella sonnei, and Escherichia coli. Infection and Immunity 40:340-350.

16. Wauters G, Vaneechoutte M. 2015. Approaches to the Identification of AerobicGram-Negative Bacteria, p 613-634, Manual of Clinical Microbiology, Eleventh Edition. American Society of Microbiology.

17. CLSI. 2008. Abbreviated Identification of Bacteria and Yeast, 2nd Edition doi:1-56238-681-6, p 64.

18. David HL, Lévy-Frébault V, Thorel M-F. 1989. Méthodes de laboratoire pour mycobactériologie clinique. Institut Pasteur.

19. Buissière J, Nardon P. 1968. Microméthode d'identification des bactéries. I. Intérèt de la quantification des caractèrs biochimiques. Ann Inst Pasteur (Paris) 115:218-231.

20. Bascomb S, Manafi M. 1998. Use of enzyme tests in characterization and identification of aerobic and facultatively anaerobic gram-positive cocci. Clinical microbiology reviews 11:318-340.

21. Pincus DH. 2006. Microbial identification using the bioMérieux Vitek® 2 system. Encyclopedia of rapid microbiological methods 1:1-32.

22. Bascomb S. 1988. 3 Enzyme Tests in Bacterial Identification, p 105-160, Methods in microbiology, vol 19. Elsevier.

23. Holding A, Collee J. 1971. Chapter I Routine biochemical tests, p 1-32, Methods in microbiology, vol 6. Elsevier.

24. Schreckenberger P, Ristow K, Krilcich A. Comparison of the Vitek Legacy, Vitek 2 Colorimetric AND Phoenix Systems for Identification of Fermenting and Non-Fermenting Bacteria of Clinical Origin, p. *In* (ed),

25. Tillotson G. 1982. An evaluation of the API-20 STREP system. Journal of clinical pathology 35:468.

26. Hall GS. 2013. Bailey & Scott's Diagnostic Microbiology, 13th Edn. American Society for Clinical Pathology.

27. Heikens E, Fleer A, Paauw A, Florijn A, Fluit A. 2005. Comparison of genotypic and phenotypic methods for species-level identification of clinical isolates of coagulase-negative staphylococci. Journal of clinical microbiology 43:2286-2290.

28. Cao Y, Wang L, Ma P, Fan W, Gu B, Ju S. 2018. Accuracy of matrix-assisted laser desorption ionization–time of flight mass spectrometry for identification of mycobacteria: a systematic review and meta-analysis. Scientific reports 8:4131.

29. Chatzigeorgiou K-S, Sergentanis TN, Tsiodras S, Hamodrakas SJ, Bagos PG. 2011. Phoenix 100 versus Vitek 2 in the identification of gram-positive and gram-negative bacteria: a comprehensive meta-analysis. Journal of clinical microbiology 49:3284-3291.

30. Posteraro B, Efremov L, Leoncini E, Amore R, Posteraro P, Ricciardi W, Sanguinetti M. 2015. Are the conventional commercial yeast identification methods still helpful in the era of new clinical microbiology diagnostics? A meta-analysis of their accuracy. Journal of clinical microbiology 53:2439-2450.

31. Posteraro B, Ruggeri A, De Carolis E, Torelli R, Vella A, De Maio F, Ricciardi W, Posteraro P, Sanguinetti M. 2013. Comparative evaluation of BD Phoenix and Vitek 2 systems for species identification of common and uncommon pathogenic yeasts. Journal of clinical microbiology 51:3841-3845.

32. Won EJ, Shin JH, Kim M-N, Choi MJ, Joo MY, Kee SJ, Shin MG, Suh SP, Ryang DW. 2014. Evaluation of the BD Phoenix system for identification of a wide spectrum of clinically important yeast species: a comparison with Vitek 2-YST. Diagnostic microbiology and infectious disease 79:477-480.

33. Grant ML, Parajuli S, Deleon-Gonsalves R, Potula R, Truant AL. 2016. Comparative evaluation of the BD phoenix yeast ID panel and Remel RapID yeast plus system for yeast identification. Canadian Journal of Infectious Diseases and Medical Microbiology 2016.

34. Singhal N, Kumar M, Kanaujia PK, Virdi JS. 2015. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. Frontiers in microbiology 6:791.

35. Novais Â, Freitas AR, Rodrigues C, Peixe L. 2019. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. European Journal of Clinical Microbiology & Infectious Diseases 38:427-448.

36. Horbach I, Naumann D, Fehrenbach FJ. 1988. Simultaneous infections with different serogroups of Legionella pneumophila investigated by routine methods

and Fourier transform infrared spectroscopy. Journal of clinical microbiology 26:1106-1110.

37. Amiali NM, Mulvey MR, Sedman J, Louie M, Simor AE, Ismail AA. 2007. Rapid identification of coagulase-negative staphylococci by Fourier transform infrared spectroscopy. Journal of microbiological methods 68:236-242.

38. Dalmasso M, Jordan K. 2015. PFGE as a Tool to Track Listeria monocytogenes in Food Processing Facilities: Case Studies, p 29-34, Pulse Field Gel Electrophoresis. Springer.

39. Pichel M, Brengi SP, Cooper KL, Ribot EM, Al-Busaidy S, Araya P, Fernández J, Vaz TI, Kam KM, Morcos M. 2012. Standardization and international multicenter validation of a PulseNet pulsed-field gel electrophoresis protocol for subtyping Shigella flexneri isolates. Foodborne pathogens and disease 9:418-424.

40. Leclair D, Pagotto F, Farber JM, Cadieux B, Austin JW. 2006. Comparison of DNA fingerprinting methods for use in investigation of type E botulism outbreaks in the Canadian Arctic. Journal of clinical microbiology 44:1635-1644.

41. Morrison D, Woodford N, Barrett S, Sisson P, Cookson B. 1999. DNA banding pattern polymorphism in vancomycin-resistant Enterococcus faecium and criteria for defining strains. Journal of clinical microbiology 37:1084-1091.

42. Lin J-J, Kuo J, Ma J. 1996. A PCR-based DNA fingerprinting technique: AFLP for molecular typing of bacteria. Nucleic acids research 24:3649-3650.

43. Clifford RJ, Milillo M, Prestwood J, Quintero R, Zurawski DV, Kwak YI, Waterman PE, Lesho EP, Mc Gann P. 2012. Detection of bacterial 16S rRNA and identification of four clinically important bacteria by real-time PCR. PloS one 7:e48558.

44. Drake RR, Boggs SR, Drake SK. 2011. Pathogen identification using mass spectrometry in the clinical microbiology laboratory. Journal of mass spectrometry 46:1223-1232.

45. Gevers D, Huys G, Swings J. 2001. Applicability of rep-PCR fingerprinting for identification of Lactobacillus species. FEMS microbiology letters 205:31-36.

46. Geha DJ, Uhl JR, Gustaferro CA, Persing DH. 1994. Multiplex PCR for identification of methicillin-resistant staphylococci in the clinical laboratory. Journal of Clinical Microbiology 32:1768-1772.

47. Breurec S, Guessennd N, Timinouni M, Le T, Cao V, Ngandjio A, Randrianirina F, Thiberge J, Kinana A, Dufougeray A. 2013. Klebsiella pneumoniae resistant to third-generation cephalosporins in five African and two Vietnamese major towns: multiclonal population structure with two major international clonal groups, CG15 and CG258. Clinical Microbiology and Infection 19:349-355.

48. Izdebski R, Baraniak A, Herda M, Fiett J, Bonten MJ, Carmeli Y, Goossens H, Hryniewicz W, Brun-Buisson C, Gniadkowski M. 2014. MLST reveals potentially high-risk international clones of Enterobacter cloacae. Journal of Antimicrobial Chemotherapy 70:48-56.

49. Juan C, Zamorano L, Mena A, Albertí S, Pérez JL, Oliver A. 2010. Metallo-β-lactamase-producing Pseudomonas putida as a reservoir of multidrug resistance elements that can be transferred to successful Pseudomonas aeruginosa clones. Journal of antimicrobial chemotherapy 65:474-478.

50. Zarrilli R, Pournaras S, Giannouli M, Tsakris A. 2013. Global evolution of multidrug-resistant Acinetobacter baumannii clonal lineages. International journal of antimicrobial agents 41:11-19.

51. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, Thomson NR. 2016. The diversity of Klebsiella pneumoniae surface polysaccharides. Microbial genomics 2.

52. Mostowy RJ, Holt KE. 2018. Diversity-generating machines: genetics of bacterial sugar-coating. Trends in microbiology 26:1008-1021.

53. Campos J, Sousa C, Mourão J, Lopes J, Pérez J, Moreno R, Antunes P, Peixe L. Assessing the potential use of matrix-assisted laser desorption ionization time-of-

flight mass spectrometry (MALDI-TOF MS) for discrimination of clinically relevant Salmonella serogroups and serotypes, p. *In* (ed),

54. Rodrigues C, Novais Â, Sousa C, Ramos H, Coque T, Cantón R, Lopes J, Peixe L. 2017. Elucidating constraints for differentiation of major human Klebsiella pneumoniae clones using MALDI-TOF MS. European Journal of Clinical Microbiology & Infectious Diseases 36:379-386.

55. Sousa C, Botelho J, Grosso F, Silva L, Lopes J, Peixe L. 2015. Unsuitability of MALDI-TOF MS to discriminate Acinetobacter baumannii clones under routine experimental conditions. Frontiers in microbiology 6:481.

56. Lartigue M-F. 2013. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for bacterial strain characterization. Infection, Genetics and Evolution 13:230-235.

57. Nikolic G. 2011. Fourier Transforms: New Analytical Approaches and FTIR Strategies. BoD–Books on Demand.

58. Santos C, Fraga ME, Kozakiewicz Z, Lima N. 2010. Fourier transform infrared as a powerful technique for the identification and characterization of filamentous fungi and yeasts. Research in microbiology 161:168-175.

59. Choo-Smith L-P, Maquelin K, Van Vreeswijk T, Bruining H, Puppels G, Thi NN, Kirschner C, Naumann D, Ami D, Villa A. 2001. Investigating microbial (micro) colony heterogeneity by vibrational spectroscopy. Appl Environ Microbiol 67:1461-1469.

60. Berthomieu C, Hienerwadel R. 2009. Fourier transform infrared (FTIR) spectroscopy. Photosynthesis Research 101:157-170.

61. Schmitt J, Flemming H-C. 1998. FTIR-spectroscopy in microbial and material analysis. 41:1-11.

62. Silhavy TJ, Kahne D, Walker S. 2010. The bacterial cell envelope. Cold Spring Harbor perspectives in biology 2:a000414.

63. Rebuffo-Scheer CA, Schmitt J, Scherer S. 2007. Differentiation of Listeria monocytogenes serovars by using artificial neural network analysis of Fourier-transformed infrared spectra. Appl Environ Microbiol 73:1036-1040.

64. Baldauf NA, Rodriguez-Romo LA, Männig A, Yousef AE, Rodriguez-Saona LE. 2007. Effect of selective growth media on the differentiation of Salmonella enterica serovars by Fourier-transform mid-infrared spectroscopy. Journal of microbiological methods 68:106-114.

65. Kim S, Reuhs B, Mauer L. 2005. Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. Journal of applied microbiology 99:411-417.

66. Kümmerle M, Scherer S, Seiler H. 1998. Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy. Appl Environ Microbiol 64:2207-2214.

67. Maquelin K, Kirschner C, Choo-Smith L-P, van den Braak N, Endtz HP, Naumann D, Puppels G. 2002. Identification of medically relevant microorganisms by vibrational spectroscopy. Journal of microbiological methods 51:255-271.

68. Beveridge TJ, Matias VR. 2006. Ultrastructure of gram-positive cell walls, p 3-11, Gram-Positive Pathogens, Second Edition. American Society of Microbiology.

69. Davis R, Mauer L. 2010. Fourier transform infrared (FT-IR) spectroscopy: a rapid tool for detection and analysis of foodborne pathogenic bacteria. Current research, technology and education topics in applied microbiology and microbial biotechnology 2:1582-1594.

70. Hancock RE. 1998. Resistance mechanisms in Pseudomonas aeruginosa and other nonfermentative gram-negative bacteria. Clinical Infectious Diseases 27:S93-S99.

71. McGowan Jr JE. 2006. Resistance in nonfermenting gram-negative bacteria: multidrug resistance to the maximum. American journal of infection control 34:S29-S37.

72. Poole K. 2001. Multidrug resistance in Gram-negative bacteria. Current opinion in microbiology 4:500-508.

73. Aguilar-Uscanga B, Francois J. 2003. A study of the yeast cell wall composition and structure in response to growth conditions and mode of cultivation. Letters in applied microbiology 37:268-274.

74. Burattini E, Cavagna M, Dell'Anna R, Campeggi FM, Monti F, Rossi F, Torriani S. 2008. A FTIR microspectroscopy study of autolysis in cells of the wine yeast Saccharomyces cerevisiae. Vibrational Spectroscopy 47:139-147.

75. Luo Y, Wang J, Liu B, Wang Z, Yuan Y, Yue T. 2015. Effect of Yeast Cell Morphology, Cell Wall Physical Structure and Chemical Composition on Patulin Adsorption. 10:e0136045.

76. Scorzoni L, De Paula E Silva ACA, Marcos CM, Assato PA, De Melo WCMA, De Oliveira HC, Costa-Orlandi CB, Mendes-Giannini MJS, Fusco-Almeida AM. 2017. Antifungal Therapy: New Advances in the Understanding and Treatment of Mycosis. 08.

77. Bosch A, Miñán A, Vescina C, Degrossi J, Gatti B, Montanaro P, Messina M, Franco M, Vay C, Schmitt J. 2008. Fourier transform infrared spectroscopy for rapid identification of nonfermenting gram-negative bacteria isolated from sputum samples from cystic fibrosis patients. Journal of clinical microbiology 46:2535-2546.

78. Curk M, Peledan F, Hubert J. 1994. Fourier transform infrared (FTIR) spectroscopy for identifying Lactobacillus species. FEMS microbiology letters 123:241-248.

79. de Sousa Marques A, de Melo MCN, Cidral TA, de Lima KMG. 2014. Feature selection strategies for identification of Staphylococcus aureus recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: a case study. Journal of microbiological methods 98:26-30.

80. Essendoubi M, Toubas D, Bouzaggou M, Pinon J-M, Manfait M, Sockalingum GD. 2005. Rapid identification of Candida species by FT-IR microspectroscopy. Biochimica et Biophysica Acta (BBA)-General Subjects 1724:239-247.

81. Lam LM, Dufresne PJ, Longtin J, Sedman J, Ismail AA. 2019. Reagent-Free Identification of Clinical Yeasts by Use of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy. Journal of clinical microbiology 57:e01739-18.

82. Lecellier A, Gaydou V, Mounier J, Hermet A, Castrec L, Barbier G, Ablain W, Manfait M, Toubas D, Sockalingum G. 2015. Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds. Food microbiology 45:126-134.

83. Mariey L, Signolle J, Amiel C, Travert J. 2001. Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. Vibrational spectroscopy 26:151-159.

84. Taha M, Hassan M, Essa S, Tartor Y. 2013. Use of Fourier transform infrared spectroscopy (FTIR) spectroscopy for rapid and accurate identification of yeasts isolated from human and animals. International journal of veterinary science and medicine 1:15-20.

85. Wenning M, Scherer S. 2013. Identification of microorganisms by FTIR spectroscopy: perspectives and limitations of the method. Applied microbiology and biotechnology 97:7111-7120.

86. Zarnowiec P, Lechowicz L, Czerwonka G, Kaca W. 2015. Fourier transform infrared spectroscopy (FTIR) as a tool for the identification and differentiation of pathogenic bacteria. Current medicinal chemistry 22:1710-1718.

87. Haag H, Gremlich H-U, Bergmann R, Sanglier J-J. 1996. Characterization and identification of actinomycetes by FT-IR spectroscopy. Journal of microbiological methods 27:157-163.

88. Carranza L. 2012. Standardization and internal validation of a bacteria identification method utilizing focal-plane-array Fourier transformed infrared spectroscopy. McGill University (Canada).

89. Ehling-Schulz M, Svensson B, Guinebretiere M-H, Lindbäck T, Andersson M, Schulz A, Fricker M, Christiansson A, Granum PE, Märtlbauer E. 2005. Emetic

toxin formation of Bacillus cereus is restricted to a single evolutionary lineage of closely related strains. Microbiology 151:183-197.

90. Lefier D, Hirst D, Holt C, Williams AG. 1997. Effect of sampling procedure and strain variation in Listeria monocytogenes on the discrimination of species in the genus Listeria by Fourier transform infrared spectroscopy and canonical variates analysis. FEMS microbiology letters 147:45-50.

91. Wenning M, Breitenwieser F, Konrad R, Huber I, Busch U, Scherer S. 2014. Identification and differentiation of food-related bacteria: a comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry. Journal of microbiological methods 103:44-52.

92. Dinkelacker AG, Vogt S, Oberhettinger P, Mauder N, Rau J, Kostrzewa M, Rossen JW, Autenrieth IB, Peter S, Liese J. 2018. Typing and species identification of clinical Klebsiella isolates by Fourier transform infrared spectroscopy and matrix-assisted laser desorption ionization–time of flight mass spectrometry. Journal of clinical microbiology 56:e00843-18.

93. Grunert T, Wenning M, Barbagelata MS, Fricker M, Sordelli DO, Buzzola FR, Ehling-Schulz M. 2013. Rapid and reliable identification of Staphylococcus aureus capsular serotypes by means of artificial neural network-assisted Fourier transform infrared spectroscopy. Journal of clinical microbiology 51:2261-2266.

94. Helm D, Labischinski H, Naumann D. 1991. Elaboration of a procedure for identification of bacteria using Fourier-transform IR spectral libraries: a stepwise correlation approach. Journal of microbiological methods 14:127-142.

95. Ghetler A. 2010. Development of an expert system for the identification of bacteria by focal plane array Fourier transform infrared spectroscopyMcGill University.

96. Toubas D, Essendoubi M, Adt I, Pinon J-M, Manfait M, Sockalingum GD. 2007. FTIR spectroscopy in medical mycology: applications to the differentiation and typing of Candida. Analytical and bioanalytical chemistry 387:1729.

97. Schmid D, Rademacher C, Kanitz EE, Frenzel E, Simons E, Allerberger F, Ehling-Schulz M. 2016. Elucidation of enterotoxigenic Bacillus cereus outbreaks in Austria

by complementary epidemiological and microbiological investigations, 2013. International journal of food microbiology 232:80-86.

98. Dieckmann R, Hammerl JA, Hahmann H, Wicke A, Kleta S, Dabrowski PW, Nitsche A, Stämmler M, Al Dahouk S, Lasch P. 2016. Rapid characterisation of Klebsiella oxytoca isolates from contaminated liquid hand soap using mass spectrometry, FTIR and Raman spectroscopy. Faraday discussions 187:353-375.

99. Johler S, Stephan R, Althaus D, Ehling-Schulz M, Grunert T. 2016. High-resolution subtyping of Staphylococcus aureus strains by means of Fourier-transform infrared spectroscopy. Systematic and applied microbiology 39:189-194.

100. Preisner O, Lopes JA, Menezes JC. 2008. Uncertainty assessment in FT-IR spectroscopy based bacteria classification models. Chemometrics and Intelligent Laboratory Systems 94:33-42.

101. Saraiva RG, Lopes JA, Machado J, Gameiro P, Feio MJ. 2014. Discrimination of single-porin Escherichia (E.) coli mutants by ATR and transmission mode FTIR spectroscopy. Journal of biophotonics 7:392-400.

102. Xie Y, Xu S, Hu Y, Chen W, He Y, Shi X. 2012. Rapid identification and classification of staphylococcus aureus by attenuated total reflectance fourier transform infrared spectroscopy. Journal of food safety 32:176-183.

103. Costa FS, Silva PP, Morais CL, Arantes TD, Milan EP, Theodoro RC, Lima KM. 2016. Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii. Analytical Methods 8:7107-7115.

104. Silva S, Tobaldini-Valerio F, Costa-de-Oliveira S, Henriques M, Azeredo J, Ferreira EC, Lopes JA, Sousa C. 2016. Discrimination of clinically relevant Candida species by Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR-ATR). RSC Advances 6:92065-92072.

105. Moreirinha C, Trindade J, Saraiva JA, Almeida A, Delgadillo I. 2018. MIR spectroscopy as alternative method for further confirmation of foodborne

pathogens Salmonella spp. and Listeria monocytogenes. Journal of food science and technology 55:3971-3978.

106. Oberreuter H, Seiler H, Scherer S. 2002. Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT-IR) spectroscopy. International journal of systematic and evolutionary microbiology 52:91-100.

107. Rebuffo CA, Schmitt J, Wenning M, von Stetten F, Scherer S. 2006. Reliable and rapid identification of Listeria monocytogenes and Listeria species by artificial neural network-based Fourier transform infrared spectroscopy. Appl Environ Microbiol 72:994-1000.

108. Rebuffo-Scheer CA, Dietrich J, Wenning M, Scherer S. 2008. Identification of five Listeria species based on infrared spectra (FTIR) using macrosamples is superior to a microsample approach. Analytical and bioanalytical chemistry 390:1629-1635.

109. Coutinho CP, Sá-Correia I, Lopes JA. 2009. Use of Fourier transform infrared spectroscopy and chemometrics to discriminate clinical isolates of bacteria of the Burkholderia cepacia complex from different species and ribopatterns. Analytical and bioanalytical chemistry 394:2161-2171.

110. Wenning M, Büchl NR, Scherer S. 2010. Species and strain identification of lactic acid bacteria using FTIR spectroscopy and artificial neural networks. Journal of biophotonics 3:493-505.

111. DeSouza L, Kamat T, Naik C. 2009. Diffuse reflectance infrared fourier transform spectroscopic (DRIFTS) investigation of E. coli, Staphylococcus aureus and Candida albicans.

112. Wohlmeister D, Vianna DRB, Helfer VE, Calil LN, Buffon A, Fuentefria AM, Corbellini VA, Pilger DA. 2017. Differentiation of Candida albicans, Candida glabrata, and Candida krusei by FT-IR and chemometrics by CHROMagar™ Candida. Journal of microbiological methods 141:121-125.

113. Sockalingum G, Sandt C, Toubas D, Gomez J, Pina P, Beguinot I, Witthuhn F, Aubert D, Allouch P, Pinon J. 2002. FTIR characterization of Candida species: a

study on some reference strains and pathogenic C. albicans isolates from HIV+ patients. Vibrational spectroscopy 28:137-146.

114. Baldauf NA, Rodriguez-Romo LA, Yousef AE, Rodriguez-Saona LE. 2006. Differentiation of selected Salmonella enterica serovars by Fourier transform mid-infrared spectroscopy. Applied spectroscopy 60:592-598.

115. Davis R, Paoli G, Mauer L. 2012. Evaluation of Fourier transform infrared (FT-IR) spectroscopy and chemometrics as a rapid approach for sub-typing Escherichia coli O157: H7 isolates. Food microbiology 31:181-190.

116. Sandt C, Sockalingum G, Aubert D, Lepan H, Lepouse C, Jaussaud M, Leon A, Pinon J, Manfait M, Toubas D. 2003. Use of Fourier-transform infrared spectroscopy for typing of Candida albicans strains isolated in intensive care units. Journal of clinical microbiology 41:954-959.

117. Martak D, Valot B, Sauget M, Cholley P, Thouverez M, Bertrand X, Hocquet D. 2019. Fourier-transform infrared spectroscopy can quickly type gram-negative bacilli responsible for hospital outbreaks. Frontiers in Microbiology 10:1440.

118. Di Gaudio F, Indelicato S, Indelicato S, Tricoli MR, Stampone G, Bongiorno D. 2018. Improvement of a rapid direct blood culture microbial identification protocol using MALDI-TOF MS and performance comparison with SepsiTyper kit. Journal of microbiological methods 155:1-7.

119. Schieffer K, Tan K, Stamper P, Somogyi A, Andrea S, Wakefield T, Romagnoli M, Chapin K, Wolk D, Carroll KC. 2014. Multicenter evaluation of the S epsityper™ extraction kit and MALDI-TOF MS for direct identification of positive blood culture isolates using the BD BACTEC™ FX and Versa TREK® diagnostic blood culture systems. Journal of applied microbiology 116:934-941.

120. Gray TJ, Thomas L, Olma T, Iredell JR, Chen SC-A. 2013. Rapid identification of Gram-negative organisms from blood culture bottles using a modified extraction method and MALDI-TOF mass spectrometry. Diagnostic microbiology and infectious disease 77:110-112.

121. Juiz P, Almela M, Melción C, Campo I, Esteban C, Pitart C, Marco F, Vila J. 2012. A comparative study of two different methods of sample preparation for positive blood cultures for the rapid identification of bacteria using MALDI-TOF MS. European journal of clinical microbiology & infectious diseases 31:1353-1358.

122. Stevenson LG, Drake SK, Murray PR. 2010. Rapid identification of bacteria in positive blood culture broths by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Journal of clinical microbiology 48:444-447.

123. Verroken A, Defourny L, Lechgar L, Magnette A, Delmée M, Glupczynski Y. 2015. Reducing time to identification of positive blood cultures with MALDI-TOF MS analysis after a 5-h subculture. European Journal of Clinical Microbiology & Infectious Diseases 34:405-413.

124. Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P-E, Rolain JM, Raoult D. 2009. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clinical Infectious Diseases 49:543-551.

125. Idelevich EA, Schüle I, Grünastel B, Wüllenweber J, Peters G, Becker K. 2014. Rapid identification of microorganisms from positive blood cultures by MALDI-TOF mass spectrometry subsequent to very short-term incubation on solid medium. Clinical microbiology and Infection 20:1001-1006.

126. Buchan BW, Riebe KM, Ledeboer NA. 2012. Comparison of the MALDI Biotyper system using Sepsityper specimen processing to routine microbiological methods for identification of bacteria from positive blood culture bottles. Journal of clinical microbiology 50:346-352.

127. Sandt C, Madoulet C, Kohler A, Allouch P, De Champs C, Manfait M, Sockalingum G. 2006. FT-IR microspectroscopy for early identification of some clinically relevant pathogens. Journal of applied microbiology 101:785-797.

Connecting statement 1

The literature review indicates the applicability of FTIR-based methods for microbial identification that have gained momentum over the past two decade, but even with the numerous successful studies, there is a lack of validation studies that support the robustness of FTIR-based methods for routine use.

Chapter 3 examines the standardization of an attenuated total reflectance (ATR) Fourier transform infrared (FTIR) spectroscopy-based method for culture, spectral acquisition, spectral preprocessing, database construction and optimization of spectral analysis algorithms to advance the field of study for potentially implementing an ATR-FTIR use in clinical microbiology and identification of foodborne microbial pathogens.

CHAPTER 3. WORKFLOW DEVELOPMENT AND STANDARDIZATION OF AN ATR-FTIR SPECTROSCOPY-BASED TECHNIQUE FOR MICROORGANISM DISCRIMINATION APPLICATIONS

3.1. ABSTRACT

Through decades of research, standardization and validation, matrix assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometry (MS) has revolutionized clinical microbiology and is cited as the first spectroscopic-based technique introduced for rapid and accurate microorganism identification. Although MALDI-TOF MS has many benefits, the initial investment of a commercial system is costly for resource-limited microbiology laboratories and has some drawbacks in differentiating between closely related microorganisms such as *Escherichia coli* and *Shigella* species. Like MALDI-TOF MS, research for the use of Fourier transform infrared (FTIR) spectroscopy in clinical microbiology requires the construction of spectral databases of well characterized microorganisms for the identification of unknown isolates and has been undertaken. Preliminary studies have demonstrated the successful use of FTIR spectroscopy in microbial identification including the discrimination between *E. coli* and *Shigella* species. Unlike MALDI-TOF MS, there is a lack of studies focusing on standardization and validation of the FTIR spectroscopy-based microorganism identification technique, limiting its implementation in routine clinical microbiology usage. The current work presented herein focuses on the development of a standard operating procedure for the use of attenuated total reflectance (ATR) mode coupled to an FTIR spectrometer for infrared spectral acquisition of microorganisms and data analysis for microorganism identification. The use of ATR-FTIR spectroscopy reduces sample preparation and simplifies spectral acquisition while yielding high quality spectral data. The spectral data are then subjected to spectral filtration, data preprocessing prior to the construction of a spectral database of well-characterized microorganisms. As part of the evaluation, the database will provide confidence limits to ensure accuracy and reliability of the results for potential future implementation in both large and resource limited clinical microbiology laboratories.

3.2. INTRODUCTION

Matrix assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) has revolutionized clinical microbiology and is the first spectroscopic technique introduced for rapid diagnostics (1). MALDI-TOF MS offers rapid microbial identification after initial culture, it can be completely automated, is low cost per sample and achieves high accuracy by comparing the peptide mass profile of an isolate to peptide mass profiles in database of well characterized microorganisms. To date, the technology is used worldwide and has been implemented for clinical routine microorganism identification of both bacteria and fungi. The implementation of MALDI-TOF MS in large clinical microbiology laboratories has reduced usage of conventional assays, leaving rapid biochemical tests for confirmation (2). Prior to MALDI-TOF MS being widely accepted and entering the clinical microbiology world for implementation, extensive work had to be carried out for its standardization, evaluation, and validation in multicentre trials. Similarly, the potential use of Fourier transform infrared (FTIR) spectroscopy-based methods for microbial identification will require standardization as the first step, followed by evaluation and validation. To date, most of the reported studies employed transmission based FTIR methods as a means of microbial identification (refer to references in Table 2.5 to 2.9). The use of attenuated total reflectance (ATR) mode for spectral acquisition however, compared to the other modes of spectral acquisition, reduces the sample preparation time greatly by not requiring the need for forming a dry film required in the use of transmission based FTIR spectroscopy studies. The current chapter and thesis focus on the development of a standardized method for the ATR-FTIR spectroscopy-based microorganism identification technique as a prelude to the development of an ATR-FTIR based infrared spectral database. The establishment of a robust spectral database must address the need for standardized culture conditions and spectral reproducibility of spectra acquired using spectrometers from different instruments manufacturers. Lastly, the chapter aims at providing a means of constructing an infrared spectral database with defined confidence limits of the isolate identification. This latter effort will require the evaluation of multiple multivariate algorithms for data mining and their use in the discrimination between isolates based solely on their infrared spectral profiles (termed spectrotyping). The aim of the chapter is to provide a comprehensive standard operating

procedure that may bring researchers and industry closer to carrying further evaluation and if warranted, validation of the ATR-FTIR based microbial identification method, paving the way for its implementation as routine in clinical microbiology laboratories.

3.3. CLINICAL MICROBIOLOGY VARIABLES FOR STANDARDIZATION

Standardization of the microorganism culture methodology and spectral acquisition parameters are extremely important for obtaining reproducible spectra of microorganisms. The following section outlines various variables investigated for standardization.

3.3.1. Sample collection method

Samples collected from hospitals or clinical reference laboratories are generally isolated from urine, sputum, positive blood cultures, pus, wounds and swabs from various parts of the body while food-related microorganisms are isolated from animal feed, contaminated food products and water. Samples can also be acquired from microbial samples stored in long-term frozen storage (stored in 10% glycerol at -80°C).

3.3.2. Sample culturing method

Frozen samples are cultured on media of choice and are subcultured once to ensure purity of the isolates, while samples collected from routine (freshly cultured from the source) are analyzed as-is. Subculturing of the routine plates may also be required to ensure purity of the isolates; this is decided by the microbiologist. The use of the 4-streak quadrant method to obtain isolated colonies and visually assess the purity of the sample in question is the current recommended practice. If the sample is impure (multiple colony morphologies; Figure 3.1), additional subculture of the sample may be undertaken. According to Clinical & Laboratory Standards Institute Guidelines (3), aerobic bacteria are cultured and incubated for 18-24 h at 35°C ± 2°C and 24-48 h at 30°C ± 2°C for yeasts (please refer to Table 3.2 for agar culture media used in this work).

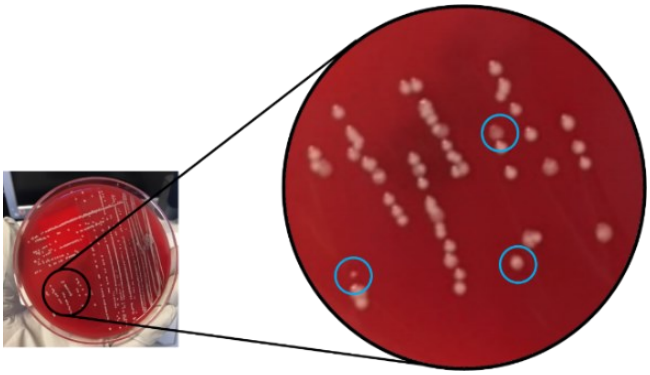


Figure 3.1. Example of mixed morphology culture on Columbia blood agar with 5% sheep blood obtained from routine to be subcultured

3.3.2.1. Effects of culture media composition on ATR-FTIR spectra of microorganisms

Different ATR-FTIR spectral profiles of the same strain can be observed as a consequence of growing the strain on different agar composition or on a specific agar produced by different manufacturers. For example, Sabouraud dextrose agar (SAB) is used while modified SAB (increased pH and reduced sugar composition) have been widely used to enhance yeast cell growth and to inhibit molds and bacteria. For the evaluation of the effects of growth media on the same strain, three pre-made SAB agar plates were acquired from (a) Difco™ Sabouraud Dextrose Agar (BD, Sparks, MD), (b) Difco™ Sabouraud Dextrose Agar Emmons (SABE) (BD, Sparks, MD) and (c) OXOID SAB (Thermo Fisher Scientific, Nepean, ON). One of the 3 agar media plates was a modified SAB plate (SABE) and the other 2 were general SAB plates (Table 3.1).

Firstly, to observe the difference in the chemical composition, the agar from each vendor was dehydrated and their infrared spectra were recorded by ATR-FTIR spectroscopy in triplicate. Triplicate ATR-FTIR spectra of each dried agar film was acquired and compared to each other by taking the average of the triplicate spectra. Figure 3.2 shows the overlaid 2nd derivative spectra and shows that the SABE spectral profile was distinct from the other 2 SAB formulations, where SABE had lower relative band intensities between 1200-900 cm^{-1} which is associated with carbohydrate absorptions. The chemical formulation of the agar (Table 3.1) provided by the manufacturer also showed SABE had 50 reduced polysaccharide content (dextrose).

Table 3.1. Composition and pH of Sabouraud dextrose agar from different manufacturer used in clinical microbiology laboratories

Media (final pH)	Manufacturer	Ingredients	(g/L)
Sabouraud Dextrose Agar (SAB) (pH 5.6 ± 0.2 @ 25°C)	Difco (BD, Sparks, MD)	Peptic digest of Animal Tissue	5
		Pancreatic digest of Casein	5
		Dextrose	40
		Agar	15
Sabouraud Dextrose Agar (SAB) (pH 5.6 ± 0.2 @ 25°C)	Oxoid (Thermo Fisher Scientific, Nepean, ON)	Mycological peptone	10
		Dextrose	40
		Agar	15
Sabouraud Dextrose Agar Emmons (SABE) (pH 6.9 ± 0.2 @ 25°C)	Difco (BD, Sparks, MD)	Peptic digest of Animal Tissue	5
		Pancreatic digest of Casein	5
		Dextrose	20
		Agar	17

Moreover, it should be noted that 2 average spectra of OXOID SAB dried agar film were obtained from 2 different lots that were manufactured a few months apart, were visually dissimilar in composition (one was more yellow in color than the other). The infrared spectra of dried agar films from the two lots were also found to be dissimilar in composition within the range of 1060-1020 cm^{-1} that may be due to the C=O stretching of carbohydrates (4-7).

The agar powder composition of Difco™ SAB and OXOID SAB are comparable, however OXOID does not define the exact “mycological peptone” components (could be animal or plant peptones in varying proportions) (Table 3.1). The undefined source and amount of peptone used in OXOID SAB agar may be attributed to the difference in color and their ATR-FTIR spectra.

The agar composition of the same culture media type (different brands) is evidently dissimilar from one another – which may have a major or minor affect on the metabolism of the microorganism. To evaluate the variability in the media composition on the spectral profiles of microorganisms, one isolate of *Candida glabrata* was cultured onto the 3 different prepared agar media using the same growing conditions (30°C ± 2°C for 48 h) before ATR-FTIR spectra were acquired in triplicates. A stacked view of the averages of the triplicate spectra of *C. glabrata* (Figure 3.2) demonstrates the spectral variability stemming from the differences in the agar composition of the same agar-type on the same strain.

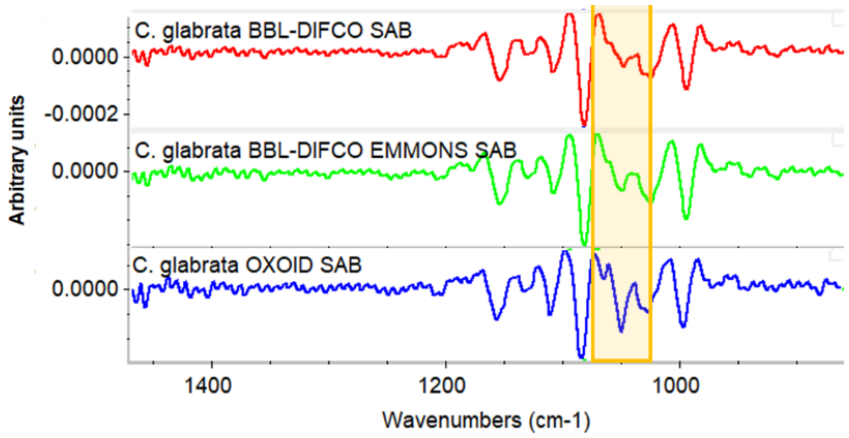


Figure 3.2. Stacked 2nd derivative of spectra of *Candida glabrata* cultured on 2 general Sabouraud dextrose agar (SAB) manufactured by Difco™ and OXOID and 1 Difco™ SAB Emmons (SABE) manufactured by Difco™

Accordingly, a standardized culture media will be recommended for optimal consistency when creating ATR-FTIR spectral databases for species identification and strain-typing of microorganisms based on FTIR spectroscopy.

Table 3.2. Selected standard growth media for culturing clinical and food microorganisms

	Clinical microbiology		Food microbiology	
	General growth media	Selective media	General growth media	Selective media
Gram-negative bacteria	BAP	MAC	BHI TSA TSAYE	XLD HEA MAC
Gram-positive bacteria	BAP	CHROMagar	BHI TSA	-
Yeasts	SAB PDA BAP	IMA CHROMagar	-	-
Abbreviations	BAP: 5% Sheep blood agar BHI: Brain heart infusion HEA: Hektoen enteric agar IMA: Inhibitory mold agar MAC: MacConkey agar		PDA: Potato dextrose agar SAB: Sabouraud dextrose agar TSA: Tryptic soy agar TSAYE: Tryptic soy agar 0.6% yeast extract XLD: Xylose Lysine Deoxycholate agar	

To further assess the effects of different culture media composition on the ATR-FTIR spectra of microorganisms, the same *Candida albicans* strain was grown on Difco™ SAB (BD, Sparks, MD) and used as reference spectra, in order to evaluate changes in the spectral profiles of the same strain grown on the following media: OXOID Columbia Blood Agar with 5% Sheep Blood (Thermo Fisher Scientific, Nepean, ON), CHROMagar™

Candida Chromogenic Media (CHROMagar, Springfield, NJ), BBL™ Inhibitory Mold Agar (BD, Sparks, MD) and BBL™ Potato Dextrose Agar (BD, Sparks, MD); now referred to as blood agar (BAP), Chrome agar (CHRM), inhibitory mold agar (IMA) and potato dextrose agar (PDA). The *C. albicans* strain grown on all culture media agar were split into 2 sets (different incubation times), set 1, incubation conditions: 30°C ± 2°C for 24 h and set 2, incubation conditions: 30°C ± 2°C for 48 h.

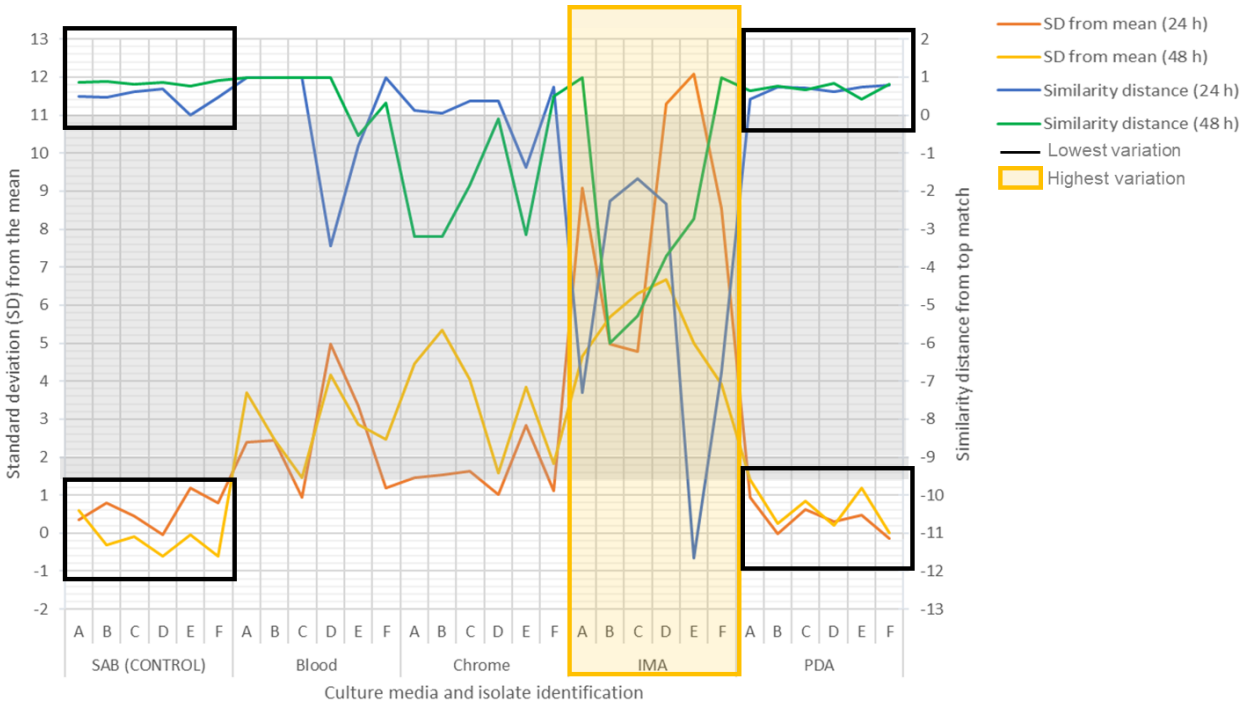


Figure 3.3. Graph illustrating the standard deviation of *Candida albicans* spectra grown on 5 different culture agar media (SAB, Blood agar, Chrome agar, IMA and PDA) from the reference *C. albicans* mean spectrum and spectral similarity distance between the test spectrum (5 replicates: A to F) and nearest reference spectrum from isolated incubated for 24 h and 48 h at 30°C ± 2°C

Per set, 5 ATR-FTIR spectra of the *C. albicans* were collected from colonies grown on the reference SAB agar plates (Difco™ SAB (BD, Sparks, MD) while *C. albicans* grown on the other plates are referred to as the test set where 5 spectra were also collected from 5 different colonies on the culture media agar plate at different incubation periods (24 h and 48 h).

To assess spectral heterogeneity between the reference spectra and test spectra, an inhouse software (developed by Dr. Andrew Ghetler; a McGill IR Group alumnus) was

utilized to calculate the standard deviation (SD) from the mean of the reference *C. albicans* and the spectral distance from the test spectrum to the nearest reference spectrum (details in later sections of the chapter). *C. albicans* grown on Difco™ SAB (BD, Sparks, MD) as the reference (main agar type used in clinical microbiology for culturing yeasts), has a SD and spectral similarity distance when incubated at 24 h comparable to growth at the 48 h mark (Figure 3.3). Isolates grown on PDA provided comparable results (Figure 3.3) at both incubation times, suggesting the composition of PDA allows for similar metabolism as *C. albicans* grown on Difco™ SAB (BD, Sparks, MD) for the *C. albicans* strain. Interestingly, upon further investigation, PDA has half the carbohydrate content compared to Difco™ SAB (BD, Sparks, MD) and does not contain any other nutrient source other than potato starch (Table A.4).

Moreover, the results obtained for IMA were highly heterogenous with extremely variable SD values from the reference mean and inadequate spectral similarity distances. This may be attributed to multiple ingredients in the agar which includes sodium chloride, manganese sulfate, sodium phosphate, magnesium sulfate and ferrous sulfate, which are used to inhibit mold growth and promote yeast growth (Table A.4). Nutrient depletion may be inconsistent during incubation causing some cells to metabolize more or faster than other cells resulting in changes in the metabolic profiles of the microorganisms yielding differences in the ATR-FTIR spectral profiles. To a lesser extent, CHRM agar exhibited spectral heterogeneity (Figure 3.3); CHRM agar is also a selective media which selects for *Candida* species.

Furthermore, Figure 3.4 demonstrates spectral heterogeneity of *Escherichia coli* (expressed as difference in the variance between replicates grown on two different agar media, BAP and MacConkey agar). *E. coli* grown on MacConkey agar (MAC) have more spectral variability in the spectral region between 1200 and 900 cm^{-1} compared to spectra of colonies grown on BAP. MAC agar is selective to lactose fermenting Gram-negative bacteria and contains significant amount of lactose (~20% of the media) compared to BAP which does not contain lactose. Accordingly, spectral variability may be attributed to variation in lactose metabolism during incubation.

As such, selecting a culture media for developing a spectral database will require the consistent use of the same culture media throughout the entirety of the work. To ensure spectral reproducibility, all microorganisms should be grown on a consistent agar media (ideally from the same media manufacturer as discussed above).

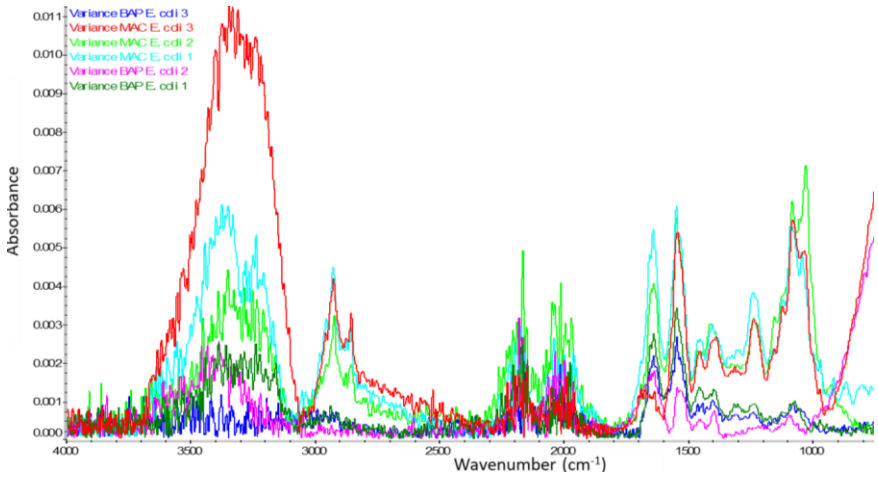


Figure 3.4. Spectral variance of triplicate spectra of three *E. coli* samples (1, 2 and 3) on BAP (dark green, dark blue and pink) and MAC agar (red, light green and light blue)

Although incubation time has little affect on the SD from the reference mean spectrum and the spectral similarity distances, not all yeasts species grow at the same rates. With the majority of clinically relevant microorganisms displaying optimal growth at 48 h, for the thesis's work, all yeast strains are grown for 48 h for consistency and are grown solely on Difco™ SAB (BD, Sparks, MD); while bacteria are grown on BAP for 24 h.

3.4. ATR-FTIR SPECTROMETERS

3.4.1. ATR-FTIR spectral deposition and acquisition parameters

After obtaining isolated colonies, 1-3 colonies are directly deposited on a 2-mm diamond using a 1 µL disposable loop. The diamond is an integral part of an ATR sampling accessory which is placed in an FTIR sampling compartment. Triplicate spectra are recorded from 3 different independent colonies acquired from the sample culture plate. Spectral acquisition parameters are pre-set where a background spectrum is recorded from a clean diamond surface prior to deposition of the microbial colonies. Upon sample deposition a spectrum of the intact colonies is immediately acquired (less than 1-2 minutes). A total of 64 co-added scans at 8 cm^{-1} resolutions over a spectral range

between 4000 and 650 cm^{-1} for the background and sample spectra. The spectrum of the sample is ratioed against the background spectrum (after Fourier transformation of the interferograms) to yield a transmittance spectrum which is converted to an absorbance spectrum. Using these parameters, an ATR-FTIR spectrum is collected in about ~1 minute per replicate, or ~3 minutes per sample (3 replicates). For traceability purposes, the spectral file name is standardized by indicating various information in the filename such as, Gram-stain type, genus, species, strain, culture media, time of incubation, date, isolate identification number, date of spectral acquisition, instrument model and operator responsible for acquiring the spectrum.

3.4.2. Performance of spectrometers

3.4.2.1. Instrument-to-instrument variability from the same vendor and instrument model

Little to no instrument-to-instrument spectral variability (of the same make and model) is essential for reproducibility of microorganism identification by the ATR-FTIR spectroscopy-based using a spectral database approach. To date, our group has reviewed multiple ATR-FTIR spectrometers of the same make and model for the Cary 630 FTIR (Agilent Technologies, CA), a Summit Pro Everest™ (Thermofisher Scientific, WI) and the Spectrum Two™ (Perkin Elmer, MA).

ATR-FTIR spectra were recorded (in triplicate) from 25 isolates of *Staphylococcus aureus* and 25 isolates of *C. albicans* on two Summit Pro Everest™ (Thermofisher Scientific, WI) ATR-FTIR spectrometers. Principal component analysis (PCA) score plots (Figure 3.5), show complete discrimination between isolates of *S. aureus* and isolates of *C. albicans* based solely on differences in the spectral absorption region between 1480 and 980 cm^{-1} using spectral data recorded on either one or both spectrometers. Using the same feature selection algorithm to attempt discrimination of the species based on the individual spectrometer proved ineffective, signifying equivalence of both spectral sets acquired from the two spectrometers. Similar results were obtained from the 2 Spectrum Two™ (Perkin Elmer, MA) spectrometers (Figure 3.6).

Unlike the two sets of spectrometers tested above, the Cary 630 ATR-FTIR (Agilent Technologies, CA) spectrometers did not perform as well when performing a region selection to ascertain if the spectral data can be assigned to a specific spectrometer. Again, using a broad region between 1480 and 980 cm^{-1} to generate the PCA plot (not shown) the *S. aureus* and *C. albicans* isolate groups can be effectively separated using either of both spectrometers. Instrument-to-instrument variability were not observed, as such, regions selection for groups of spectra, instrument-to-instrument variability was also not observed.

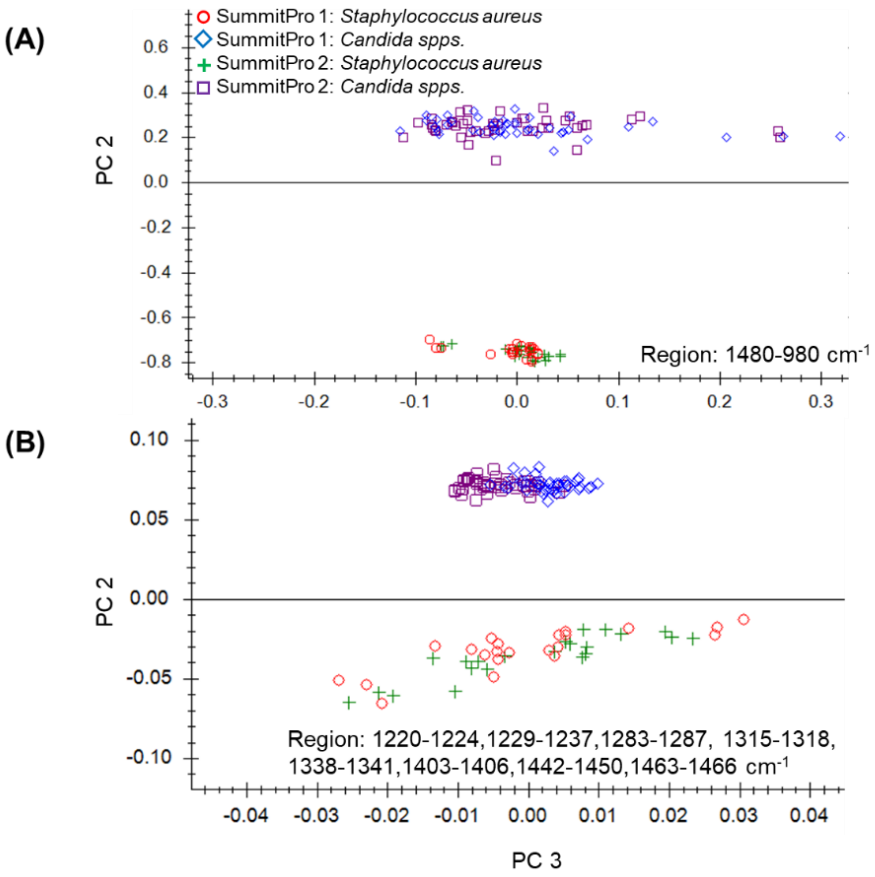


Figure 3.5. Principal component score plot of same strain of *Staphylococcus aureus* and *Candida* spp. collected on 2 spectrometers (Summit Pro Everest™ (Thermofisher Scientific, WI)) to evaluate spectral variability between instruments where (A) is broad region and (B) is generated with regions selected with a feature selection algorithm to delineate the lack of significant differences between the two spectrometers

For microorganism identification applications, instrument-to-instrument evaluations should be performed to assess spectral reproducibility to ensure reproducible results between different microbiology laboratories using the same spectral reference databases.

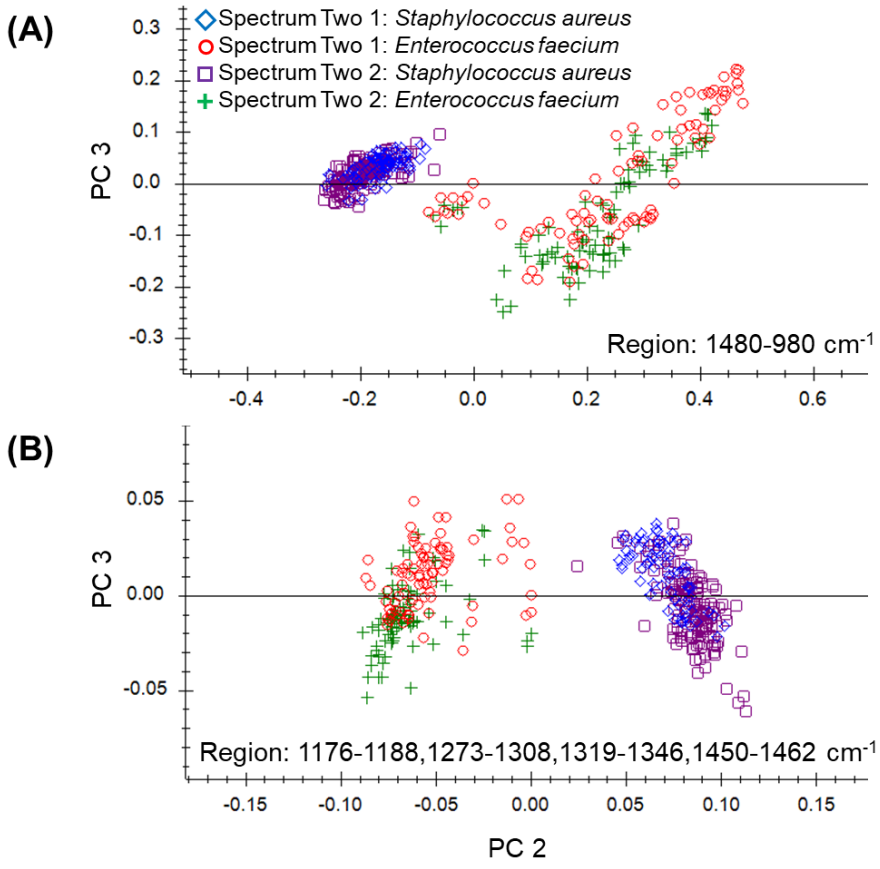


Figure 3.6. Principal component score plot of same strain of *Staphylococcus aureus* and *Enterococcus faecium*. collected on 2 spectrometers (Spectrometer Two™ (Perkin Elmer, MA)) of the same make and model to evaluate spectral variability between instruments where (A) is broad region and (B) is generated with regions selected with a feature selection algorithm between the two different spectrometers

3.4.2.2. Different make and model instrument-to-instrument variability

There are multiple manufacturers of FTIR spectrometers, each with their own unique optical and electronic components such as the interferometer, detector, amplifier, and analog-to-digital leading to differences in spectral range coverage, spectral resolution and signal-to-noise ratios (SNR). With advancement of modern technology, FTIR spectrometers have been reduced in size while maintaining or improving SNR of the instrument.

The SNR is measured without any sampling module (or accessory) and therefore provides a spectrum when the highest energy reaching the detector affording the optimal SNR as a defined spectral measuring time (e.g., 30 seconds spectral acquisition time or to the number of co-added scans, 64 scans), and at a specified spectral resolution (e.g., 8 cm^{-1}). Placing an ATR sampling accessory in the optical path can result in substantial drop in energy throughput (>50% reduction) with the consequence of reducing the SNR. In order to improve the SNR, increasing the spectral measurement time (increasing number of co-added scans) will be needed at the same specified spectral resolution.

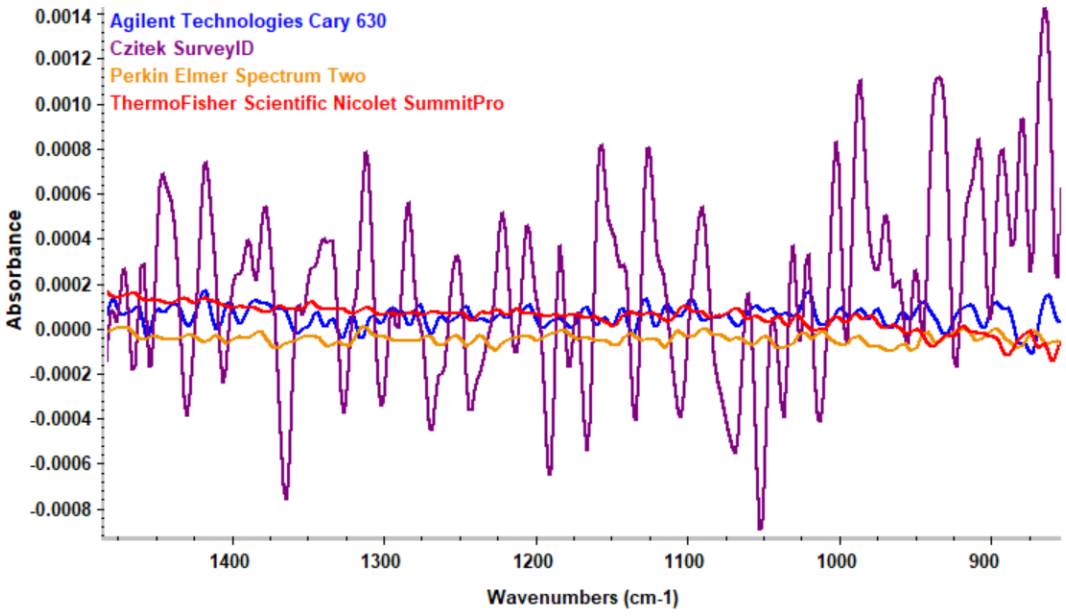


Figure 3.7. Overlay spectra of 100% lines acquired from 4 different spectrometers to demonstrate the relative SNR in the infrared spectral absorption region between 1480 and 880 cm^{-1}. Spectra were collected with 64 co-added background scan and 64 co-added sample scans in the absence of sample at 8 cm^{-1} spectral resolution. It should be noted that the lower SNR of the Czitek SurveyID is an integrated FTIR microscope with a more limited energy throughput.

To date, our group has assessed multiple ATR-FTIR spectrometers (same model) from four different manufacturers [Cary 630 FTIR (Agilent Technologies, CA), a Summit Pro Everest™ (Thermofisher Scientific, WI), SurveyIR™ (Czitek, CT) and a Spectrum Two™ (Perkin Elmer, MA)]. Observing the residual noise between the 4 spectrometers (Figure 3.7 and Figure 3.8), there are clear sensitivity differences between the spectrometers. In order of increasing sensitivity, SurveyIR™ < Cary 630 FTIR < SummitPro Everest™ <

Spectrometer Two™. Considering the Czitek SurveyIR™ spectrometer has a removable ATR accessory (has to be removed to deposit the microbial sample and replaced on repeatedly per sample), the variance between triplicate replicate spectra of one sample is high compared to the other spectrometers (Figure 3.9), it is therefore not suited for the application of microorganism identification and should not be considered at this time for use.

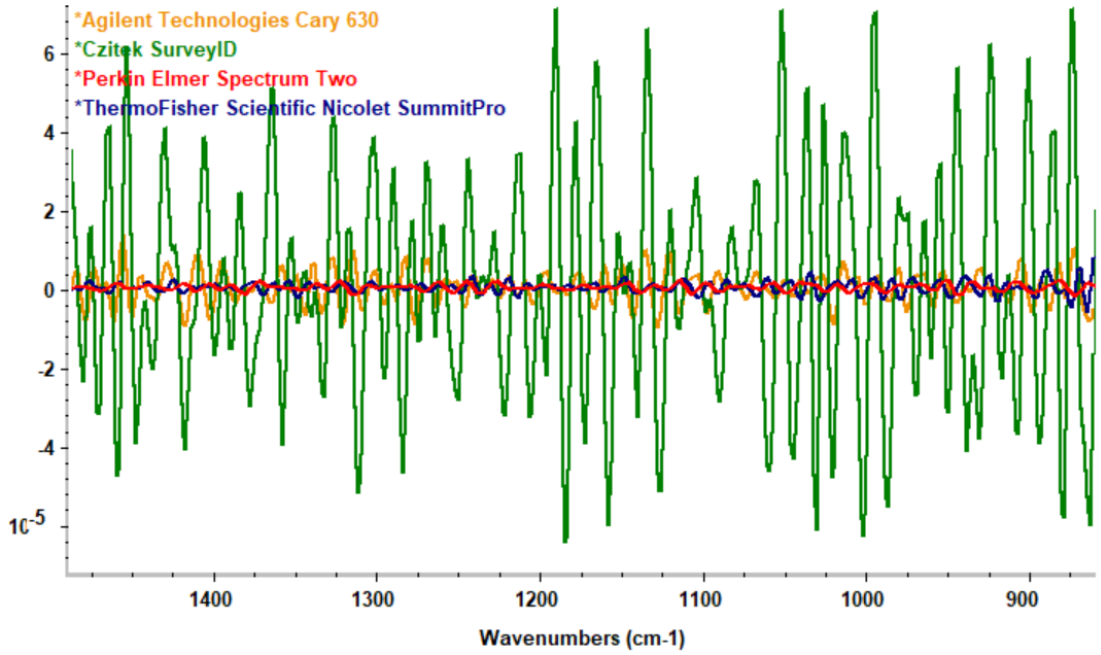


Figure 3.8. Overlay of 2nd derivative spectra obtained from 4 different spectrometers to demonstrate noise level in the infrared spectral absorption region of 1480 and 880 cm^{-1}. Spectra were collected with 64 co-added background scan and 64 co-added sample scans in the absence of sample at 8 cm^{-1} resolution.

The effects of different spectrometer models may contribute to the overall performance of the ATR-FTIR spectroscopy-based microorganism identification method, such that a spectral reference database may consist of spectra belonging to one model, while spectra being collected for microbial identification are acquired from multiple models (and manufacturers). If the spectral reference database is created without the representation of all make and models and an end-user collect spectra for microorganism identification from one that is not represented in the spectral reference database, then the performance

of the microbial identification technique will be reduced compared to if there were representation of all possible spectrometers being used.

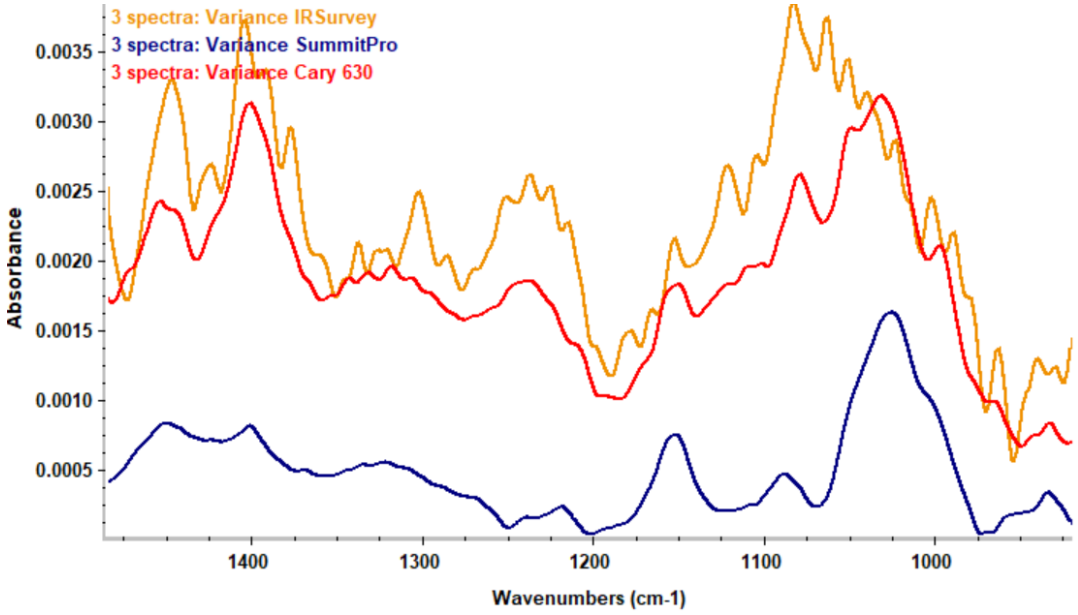


Figure 3.9. Overlay variance of triplicate replicate spectra of the same-strain *Staphylococcus aureus* (MRSA NS051) on three different ATR-FTIR spectrometers: Czitek IRSurvey, Thermofisher Summit Pro and Agilent Technologies Cary 630

For example, ATR-FTIR spectra (in triplicate) of 50 different strains of Gram-positive bacteria belonging to *S. aureus*, *Enterococcus faecalis* and *Enterococcus faecium*, were collected (using the same spectral acquisition and preprocessing parameters) using 3 different spectrometers acquired from 3 different FTIR manufacturers; Cary 630 FTIR (Agilent Technologies, CA), a Summit Pro Everest™ (Thermofisher Scientific, WI), and a Spectrum Two™ (Perkin Elmer, MA). The PCA score plots were generated from the differences in the spectral profiles of the microorganisms in the region between 1480 and 980 cm^{-1}, show that *S. aureus*, *E. faecalis* and *E. faecium* can be separated effectively from each other (Figure 3.10). Furthermore, the PCA score plot also reveals the spectral equivalence between spectra recorded on the Summit Pro Everest™ (Thermofisher Scientific, WI), and a Spectrum Two™ (Perkin Elmer, MA) versus the spectral group recorded on the Cary 630 FTIR spectrometer (Agilent Technologies, CA), (Figure 3.10, dashed line). In addition, spectra of *S. aureus* acquired using the Agilent ATR-FTIR demonstrates more spectral variability within the species compared to the other two

spectrometers. The use of region selection algorithms, regions that contribute to spectrometer variability can be omitted. Figure 3.11 shows the discrimination between *S. aureus*, *E. faecalis* and *E. faecium* is possible without the observed separation between the 3 systems shown in Figure 3.10 using a broad spectral region. Accordingly, a database that is independent of instrument manufacturer maybe possible with using a limited spectral region(s).

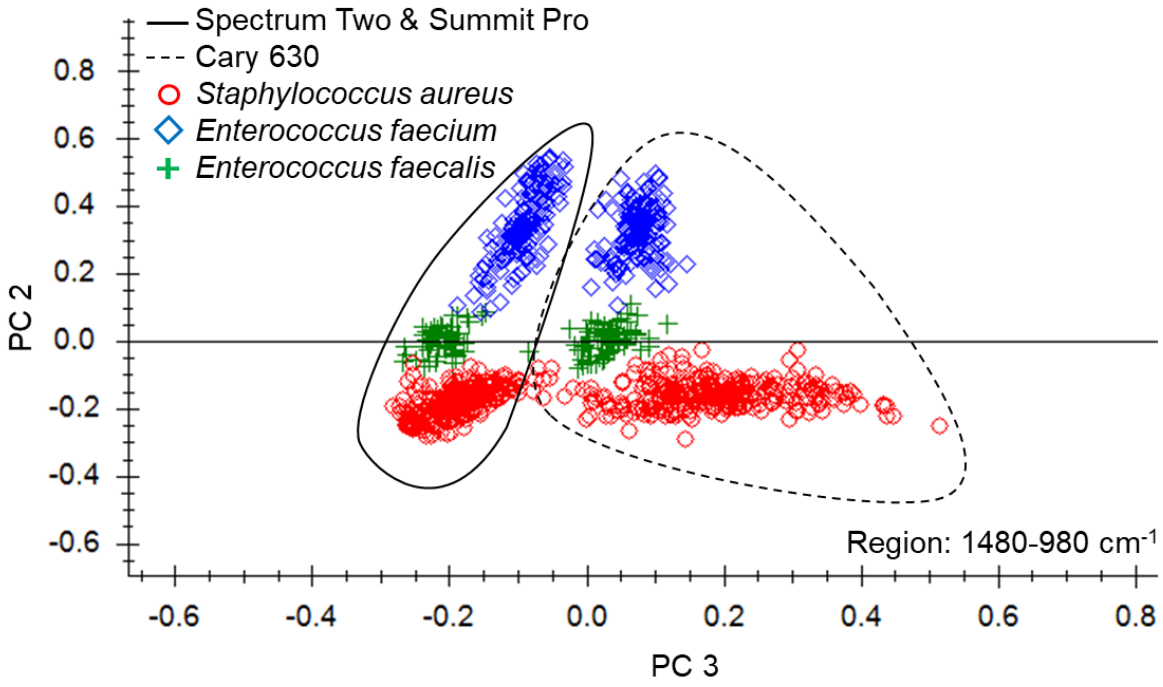


Figure 3.10. Principal component analysis score plot (PC 3 and PC 2) of 50 strains of *Staphylococcus aureus*, *Enterococcus faecium* and *Enterococcus faecalis* each on 3 different spectrometers (Cary 630 FTIR (Agilent Technologies, CA), a SummitPro Everest™ (Thermofisher Scientific, WI), and a Spectrometer Two™ (Perkin Elmer, MA))

This is an extremely important note to consider for when developing spectral reference databases for the intent of microorganism identification using the current outlined method. Creating reference spectral database using one spectrometer model will result in poor microorganism identification performance if a different spectrometer model is used to collect routine isolates for spectral database interrogation. Only spectrometer models used to collect spectra for the construction of the spectral reference database may be used for microbial identification.

Using region selection algorithms, the data collected from the 3 spectrometers may be combined to generate a spectrometer independent database. In the current work, due to availability of multiple spectrometers from one model at the time of the research, only the Cary 630 FTIR (Agilent Technologies, CA) was used in the following chapters or as stated otherwise.

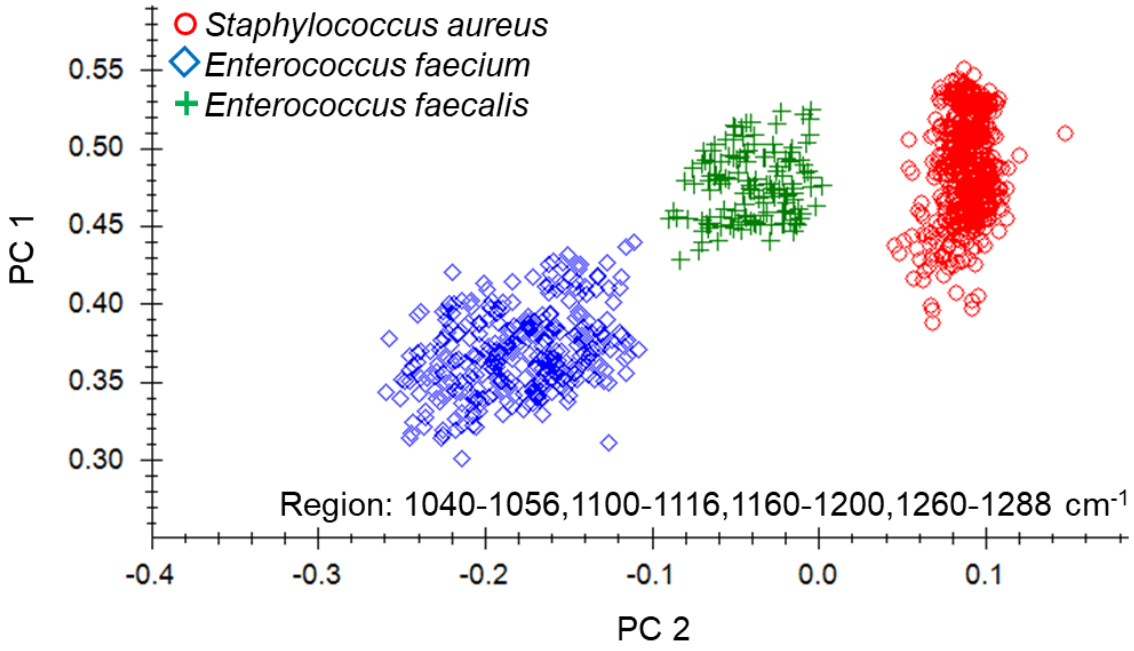


Figure 3.11. Principal component analysis score plot (PC 2 verses PC 1) of 50 strains of *Staphylococcus aureus*, *Enterococcus faecium* and *Enterococcus faecalis* recorded on 3 different spectrometers (Cary 630 FTIR (Agilent Technologies, CA), a Summit Pro Everest™ (Thermofisher Scientific, WI), and a Spectrum Two™ (Perkin Elmer, MA)). The spectral regions for the PCA plot spectral regions identified with the use of a feature selection algorithm for the discrimination between the three species.

3.5. EFFECTS OF SAMPLE DEPOSITION PROTOCOL ON ATR-FTIR SPECTRAL REPRODUCIBILITY

The spectra of microorganisms cultured on agar plates are acquired by placing 1-2 isolated intact colonies onto the diamond surface of an ATR-FTIR sampling surface. No sample preparation after incubation is needed and no extraction step is required; the outlined method is a direct colony analysis method.

3.5.1. Sample moisture control as a criterion for achieving reproducible spectra

Direct colony analysis from culture agar plates without drying the sample to create a film will require standardization of moisture content. Based on our studies, we have observed that moisture loss/retention differs from genus to genus. Accordingly, spectral reproducibility of microorganisms can be affected by rapid water evaporation during data collection or upon deposition onto the ATR sampling surface. The water has strong absorptions in the infrared spectrum, varying moisture content of the microorganism will be observed in the spectra. Beyond the spectral contribution of water, the biochemical components, such as proteins and polysaccharides participate in creating an H-bond network that stabilizes superstructure of microorganisms, such as the cell wall, ribosomes, cytoplasm and other components. Desiccation (or partial desiccation) of the microorganism will result in changes in band intensities and bandwidths in the spectra (Figure 3.13).

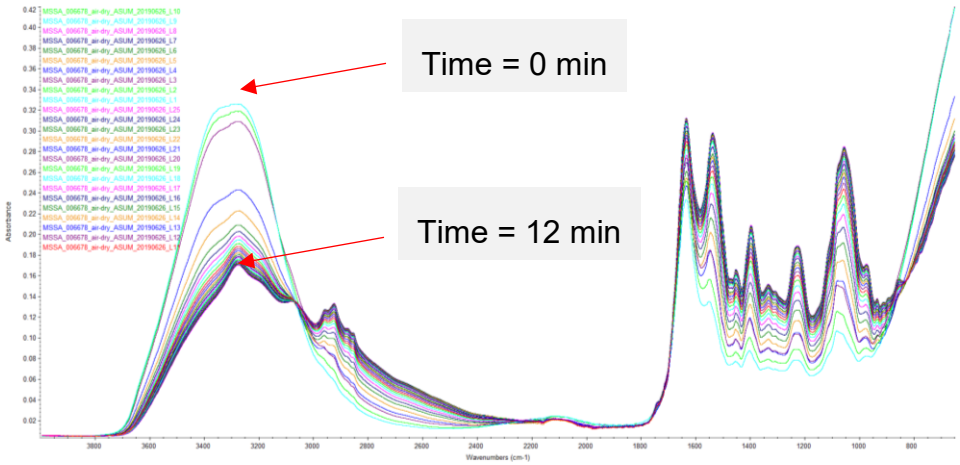


Figure 3.12. Twenty-five spectra of *Staphylococcus aureu*s over a time frame of 12 minutes (30 seconds between each spectral acquisition)

For instance, Figure 3.12 illustrates 25 superimposed spectra (32 scans each) of 1 sample (*S. aureus*) drying over a 12.5-minute period (30 seconds between each scan) at room temperature of the surface on an ATR sampling surface. After the first scan (30 seconds), the sample began to dry and by ~1.5 minutes from the initial sample deposition, there was a significant loss in moisture. Spectral variability between the spectra of the

same species (e.g., different strains of *S. aureus*) may affect the efficacy of microorganism identification by the ATR-FTIR spectroscopy-based method.

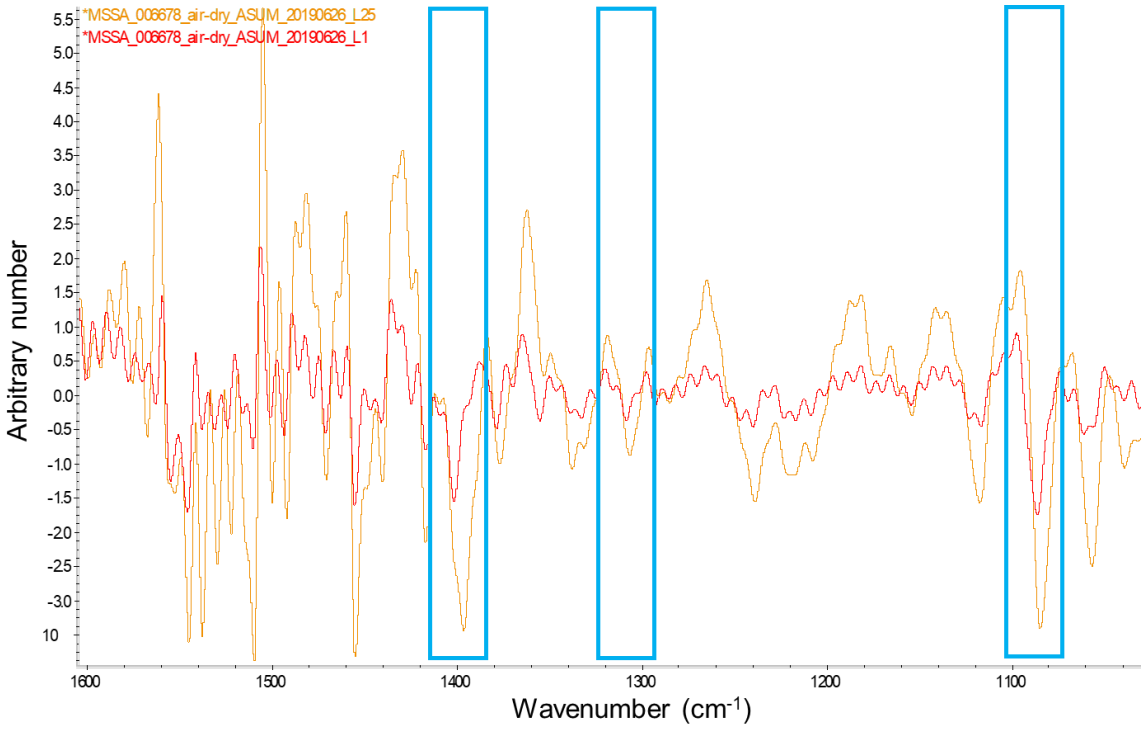


Figure 3.13. Superimposed second derivative ATR-FTIR spectra of *Staphylococcus aureus* of wet colonies (red) and air-dried colonies (orange) illustrating peak shifts between 1800-1000 cm^{-1}

To overcome rapid moisture loss, an agar cap was developed to enclose the freshly deposited colonies on the ATR sampling surface. The water activity (A_w) of the agar in the agar cap in water was obtained at room temperature (RT) for 6 different formulations (Table 3.3). As a result, the formulation of 1% agar in water with a water activity of 0.999 was selected first for the experiment and the formulation with 0.5% agar was rejected due to being too wet and unstable.

Using the selected formulation, three 1% agar in water caps were made with 3 different headspace volumes, 1060.29 mm^3, 176.71 mm^3 and 0 mm^3 or, 6-mm, 1-mm and 0-mm (agar in contact with the colony) distance from the ATR sampling surface to the surface of the agar of the agar cap, respectively. Evaluating the difference in headspace volumes for the formulation, using the same *S. aureus* strain as above, a total of 25 spectra (32

scans each) were collected over a 12.5-minute period at RT. The collected spectra (*n*=75) were compared to reference *S. aureus* spectra (*n*=100) previously collected immediately after sample deposition (Figure 3.14).

Table 3.3. Water activity of different agar and salt composition for the development of a microbial colony moisture control cap

Water activity at RT	Agar composition (%)	NaCl composition (%)
1.000	0.5	0
0.999	1	0
0.996	1.5	0
0.977	2	0
0.972	1.5	7.5
0.956	2	7.5

From the plotted area in (Figure 3.14), any SD value from the reference mean below the shaded area is significant and spectral similarity distances above the shaded area are significant – any value within the shaded area is insignificant, relative to the reference spectra. As observed in Figure 3.14, the largest volume of headspace is the least performant, and the spectral quality diminishes after 30 seconds (sample started to dry on the ATR-FTIR sampling surface). The headspace humidity of 6-mm distance from the ATR sampling surface to the agar surface of the agar cap is insufficient to maintain colony moisture or maintain the spectral quality over time (soon after sample deposition). With the cap with the 1-mm distance from the ATR sampling surface to agar surface, minimal spectral quality was lost, and spectral acquisition after 4 minutes was still consistent (comparable) to the spectra acquired immediately after sample deposition. Lastly, with zero headspace volume, the least spectral variation was observed but due to safety concerns (microorganism transfer to the agar cap), the design is not practical, and hydration of the cells may weaken absorption bands that may be vital for species identification or strain-typing. The 1% agar in water (0.99 A_w) cap design with a 176.71 mm^3 headspace volume (or 1-mm distance) between the ATR sampling surface and the agar surface is therefore an optimal design in maintaining moisture content for fast-drying microbial colonies (Figure A.3).

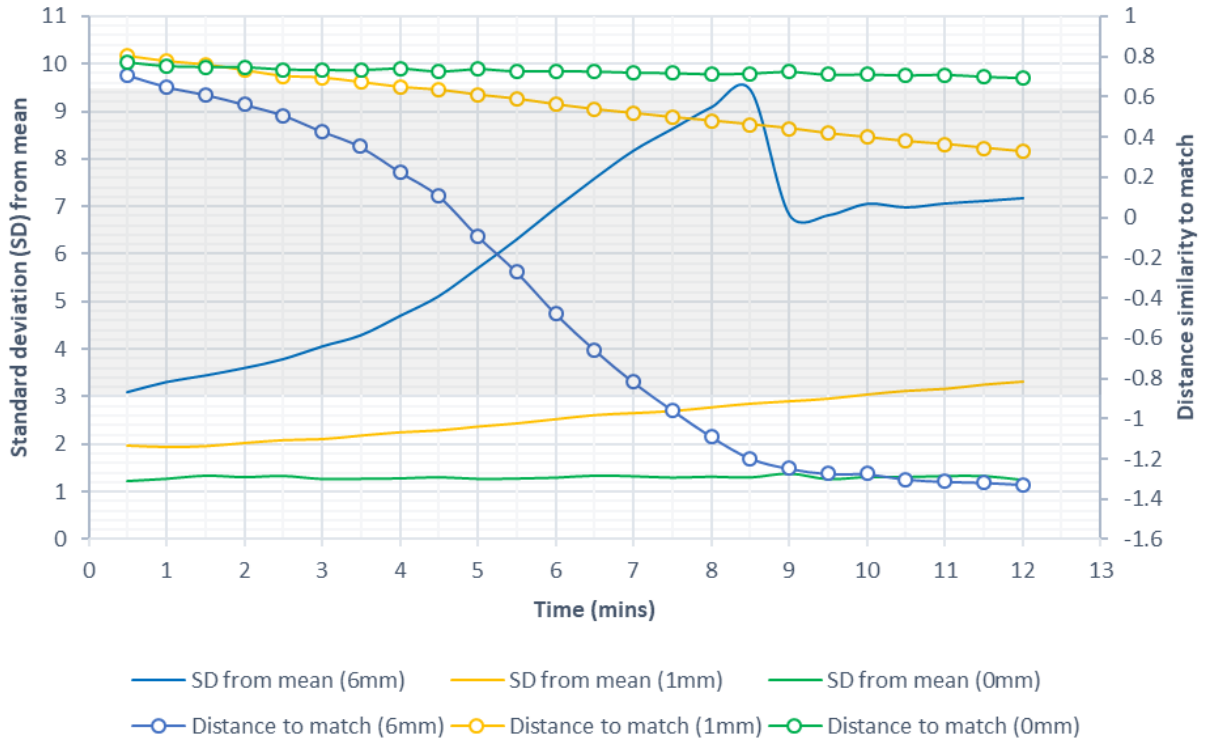


Figure 3.14. Evaluation of the effects of headspace volume humidity (simplified by height between agar and sampling surface) from 1% agar in water on the spectra of *S. aureus* as a function of time. The standard deviation from the mean and similarity distance is relative to a set of *S. aureus* spectra previously collected after immediate deposition of the sample

Multiple studies have followed the sample drying methods for microorganism identification by transmission FTIR spectroscopy (references in Table 2.5, Table 2.6, Table 2.7, Table 2.8 and Table 2.9); the current study is focused on standardizing ATR-FTIR spectroscopy from direct culture agar plate without any sample processing steps (wet analysis, no drying). With ATR-FTIR spectra being highly affected by microbial hydration, it is important to standardize the moisture content of microorganisms on the ATR sampling surface to maintain consistency throughout analysis. As drying samples on an ATR sampling surface requires additional time (varies from 5 to 30 minutes for complete drying), it would not be practical for implementation in high volume microbiology laboratories. As such, the developed agar cap would maintain moisture for fast-drying microorganisms (e.g., *Enterococcus* species and *Streptococcus* species) however, it may be used throughout all spectral acquisition to maintain spectral consistency.

3.6. DATA ANALYSIS

Our lab group have previously created an in-house software DataAnalysis (McGill University, QC) which is used to process and analyze FTIR spectra with feature selection, PCA and hierarchical cluster analysis (HCA). Dr. Andrew Ghetler (graduated from the McGill IR group) created the software to include a feature selection algorithm which performs a grid-greedy search of the spectral features to determine the spectral regions which contribute to the highest discrimination between defined spectral groups (or classes). As such, the software is used for both supervised and unsupervised analysis. Besides using a broad spectral region for HCA and PCA (unsupervised analysis), the obtained regions from a feature selection can be used with HCA or PCA and would therefore be considered a supervised analysis.

In addition to the in-house software, JMP® Pro 15 (SAS Institute Inc, NC) is also utilized for data analysis. JMP® Pro 15 (SAS Institute Inc, NC) is a general statistical software with multiple unique features such as creating visual representation of datasets, cluster analysis, multivariate analysis while various machine learning algorithms are also available such as neural net and support vector machines (SVM) analysis.

Data analysis has been optimized using the software mentioned above and are used throughout the thesis and current sub-section.

3.6.1. Spectral filtration and pre-processing

Through DataAnalysis (McGill University, QC), spectral quality (SQ) checks are completed through visual inspection of the infrared spectral profiles or, by using peak height measurement of selected peaks (e.g., the magnitude of the water absorption band at ~3400 cm^{-1} or amide II band (~1550 cm^{-1}) or PO_2^- (~1080 cm^{-1}) absorption bands in the infrared spectrum). The spectral filtration process is aimed at removing spectra of microorganisms displaying low biomass (based on the low absorbance in the lipopolysaccharide absorption region ~1085 cm^{-1}) or a weak amide II band intensity, or loss of moisture samples (due to partial inadvertent sample drying) assessed by low absorption of the water band centered at 3333 cm^{-1}). HCA and PCA can also be an

effective process to detect outliers (extreme spectral distances away from most of the spectra).

If any spectrum fails at least one of the 3 SQ checks, then the spectrum will be removed from the analyses as it does not meet the spectral quality criteria. The three SQ checks are described in detail below and must be followed when constructing spectral reference databases and/or during real-time analysis.

SQ check 1:

After calculating the absorption differences between the peak height (center of major water absorption band at 3333 cm^{-1}) and a single baseline point at 1633 cm^{-1} (amide I absorption band that is highly affected by water absorption), any spectra outside the peak height limits of [0.05-0.22] absorbance intensity from a baseline point of 2500 cm^{-1} are eliminated as they signify loss of water content from the microbial cells (i.e., the sample has partially or completely dried). Figure 3.15 demonstrates the SQ1 check with two spectra of the same bacterial sample with different water content where the first spectrum is acceptable and the second is not.

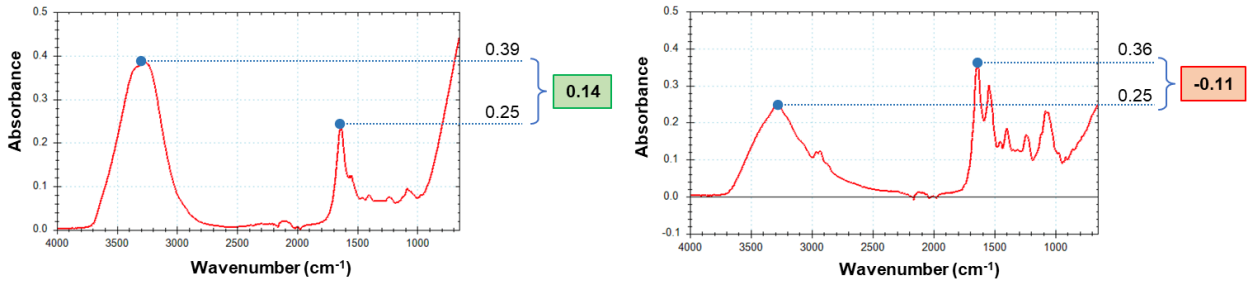


Figure 3.15. Spectral Quality check 1 example of a good spectrum with a peak height difference within the range of [0.05-0.22] absorbance intensity from a baseline point of 2500 cm^{-1} for peaks at 3333 cm^{-1} (ν(O-H)) and 1633 cm^{-1} (ν(amide I and H-O-H bend)) and a bad spectrum that does not fall in peak height range in those peaks

SQ check 2:

After calculating the absorption differences between the peak height at 1085 cm^{-1} and a single baseline point at 1200 cm^{-1}, any spectra falling outside of the peak height limits of [0.013-0.07] are also omitted as they signify the absence of significant biomass in the sample.

SQ check 3:

Calculating the peak area between 1480-980 cm^{-1} (baseline correction covering the region between 1800 and 980 cm^{-1}), all samples not within the limits of [4.0-21] will also be eliminated due to low biomass of the sample.

After spectral filtration, the 1st derivative and vector normalization at 1480-980 cm^{-1} are calculated for the remaining spectra. Although taking the first derivative of a spectrum results in an increase in spectral noise, it removes any baseline offsets or tilts in addition to enhancing the peak separation. Vector normalization of the first derivative is taken to compensate for remaining variabilities (biomass and to a lesser extent water content) in the filtered spectral set.

3.6.2. ATR-FTIR pairwise spectrotyping of microorganisms: a first step in the construction of a spectral database for microorganism identification

Spectrotyping is a word coined by the McGill IR Group, which is defined as the process of determining the differences between spectra of microorganisms based on the absence or presence, and relative intensities of infrared absorption bands. Spectra with similar absorption bands and relative intensities are therefore classified as a spectrotype. The words spectrotyping and spectrotype will be used in the current and following chapters of the thesis.

Two microorganism identification methods based on the use of ATR-FTIR spectroscopy are presented in the following section. Both methods are based on constructing a decision tree-like (or sequential multitier) spectral search databases (often used interchangeably with “library”) where method (I) is based on microorganism identification by selecting spectral features (spectrotyping) and k-nearest neighbor(s) (k-NN) while the second method (II) is based on the use of SVM learning algorithm for microbial classification. Two different techniques evaluated are meant to act as a stand-alone or combined method for microorganism identification depending on the spectral database structure, to be discussed in the following subsections.

3.6.2.1. I: Application of multivariate statistical analysis for the spectrotyping approach of microorganism identification

Spectrotyping as a method for microorganism identification consists of 2 parts, (i) determining spectral regions specific to two groups (or classes) of spectra (e.g., between families, genera, species, strain-type, serogroup and/or spectrally similar groups; *E. coli* versus *Shigella* species) through the application of HCA, PCA in combination with the feature selection algorithm; and (ii) database construction based on findings of part (i) and classification by k-NN. PCA and HCA are used to visualize relationships between clusters. PCA is a method that is used to reduce dimensionality of large datasets such as FTIR spectral data by transforming spectral data into smaller sets of variables (principal components or PC) to reduce computation time and reduce redundant information. The highest percentage of explained variances in the dataset are found in PC 1 and reduces with PCs 2, 3 and 4 and so on. PCA is often visualized by a PC score plot and will be presented in the current section.

HCA on the other hand is a clustering method that groups similar objects (or spectra) together. In terms of a spectral dataset, HCA starts with identifying 2 of the closest spectra (spectral similarity through cosine similarity distance metric, or other distance metric of choice) and iteratively continue to cluster the similar spectra and/or merge similar clusters until no more spectra or clusters are available. The data can be presented in terms of a distance matrix or a dendrogram, of which, the latter is the visual representation commonly used for HCA. In our work we have opted to utilize the cosine similarity distance and ward linkage in the generation of the dendrogram.

Both PCA and HCA can be computed using unsupervised methods based on the use of broad spectral regions or supervised where specific regions are identified using a feature selection algorithm to differentiate between two or more user defined classes, followed by the use of selected datapoints to generate the PCA and HCA plots. The feature selection algorithm (termed grid-greedy search) can reduce the number of spectral datapoints significantly. The grid search starts by searching wavenumbers in increments of 6-20 cm^{-1} over a user-defined spectral range(s) to produce sets of wavenumbers that maximized the spectral differences between two or more user-defined classes using k-

NN, achieving the narrowest spectral region with the highest classification score (CS) between the sets of regions from the grid search. Where CS is calculated as CS – N/k where N is the number of strains (spectra), and k is the k-NN. Further details are described elsewhere (8, 9).

3.6.2.1.1. Example of spectrotyping methodology for spectral database construction

HCA and PCA plots generated using defined spectral regions that optimize the separation of two distinct classes are employed in the development of a multitier decision-like tree structure for classification of microorganisms. For example, a pool of spectra belonging to 4 *Staphylococcus* species: *S. aureus*, *S. hominis*, *S. epidermidis* and *S. lugdunensis*; the latter 3 species can be grouped into one group (classified as coagulase negative staphylococci (CoNS)) while, the second group can be assigned to *S. aureus* (a coagulase-positive staphylococci). The first implicit assumption is that the CoNS class will differ in its biochemical profile from the biochemical profile of *S. aureus.* Furthermore, these differences are more significant than differences between CoNS species. The second assumption is that the biochemical changes are reflected in the spectral differences between the two groups (CoNS versus *S. aureus)*. Unless these two conditions are met, the use of infrared spectroscopy as a technique for the identification of microorganisms would not be possible. To test this hypothesis, the two spectral data sets (belonging to *S. aureus* and CoNS groups) were used to generate PCA and HCA plots using the fingerprint region of their respective infrared spectra. Figure 3.16 shows PCA plots generated using different spectral regions of the two groups illustrating an effective separation between the two groups. Similarly, Figure 3.17 illustrates the HCA plots also showing that the same spectral region can generate two principal and distinct clusters belonging to *S. aureus* and CoNS. It is also of interest to note that HCA was also effective at illustrating the discrimination between the three species belonging to the CoNS group (*S. hominis*, *S. epidermidis* and *S. lugdunensis*). The selected regions, 1023-1034, 1060-1066, 1070-1075, 1329-1334, 1349-1355 cm^{-1}, were the best for complete discrimination (Figure 3.16 (B) and Figure 3.17, (B)). The 2 classes CoNS and *S. aureus* are then considered as a pair where the next pair would be searched for in the CoNS

group of spectra while *S. aureus* is a single species and is the final group for microorganism identification.

In the given example, the CoNS group has 3 species in which, using the same approach, the second pair of groups of spectra are *S. epidermidis/S. hominis* and *S. lugdunensis* observed using spectral regions 1085-1090, 1092-1101, 1120-1128, 1318-1323 cm^{-1}. The third set (consiting of two pairs) is *S. epidermidis* and *S. hominis* which can be differentiated from each other using the spectral regions 1032-1040, 1042-1047, 1241-1247, 1277-1284, 1321-1327, 1446-1454 cm^{-1} (Figure 3.16 and Figure 3.17, PC score plots (C) and (D)). The three sets are then stringed together to create the decision tree-like spectral database structure (Figure 3.18).

It should be noted that this is a spectrotyping technique and selection of pairs are not necessarily reflective of current taxa of microorganisms. Current taxonomy classification aids in pre-selecting the pairs, however, there are times where the microoroganism's spectral fingerprint does not relate to its current taxa. Classification of microorganisms is influenced by phenotype and today, are moving towards classification by genetic relationship based on previous ancestors. An FTIR spectrum is a snapshot of the current biochemical composition and metabolomic status. As such, using a feature selection algorithm may descibe more of the differences and similarities between the pairs (relative to carbohydrate, protein and lipid composition).

Once the spectral pairs are strung together as depicted in Figure 3.18, it is interesting to note the increase in spectral similarity between pairs with increasing levels (based on cosine distances). For example, the spectral similarity between CoNS and *S. aureus* is smaller than the spectral similarity between *S. epidermidis* and *S. hominis*. When developing the multitier spectral database, considering that taxa of microorganisms may be a thing of the past when developing identification models using spectroscopic techniques, it can be used as a template for developing the structure. As such, datamining can be greatly simplified by classifying microorganisms by their cell wall composition (O-antigen, H-antigen, polysaccharides, lipids, and proteins) yielding a spectrally based classification method commensurate with the biochemical composition of the microorganism.

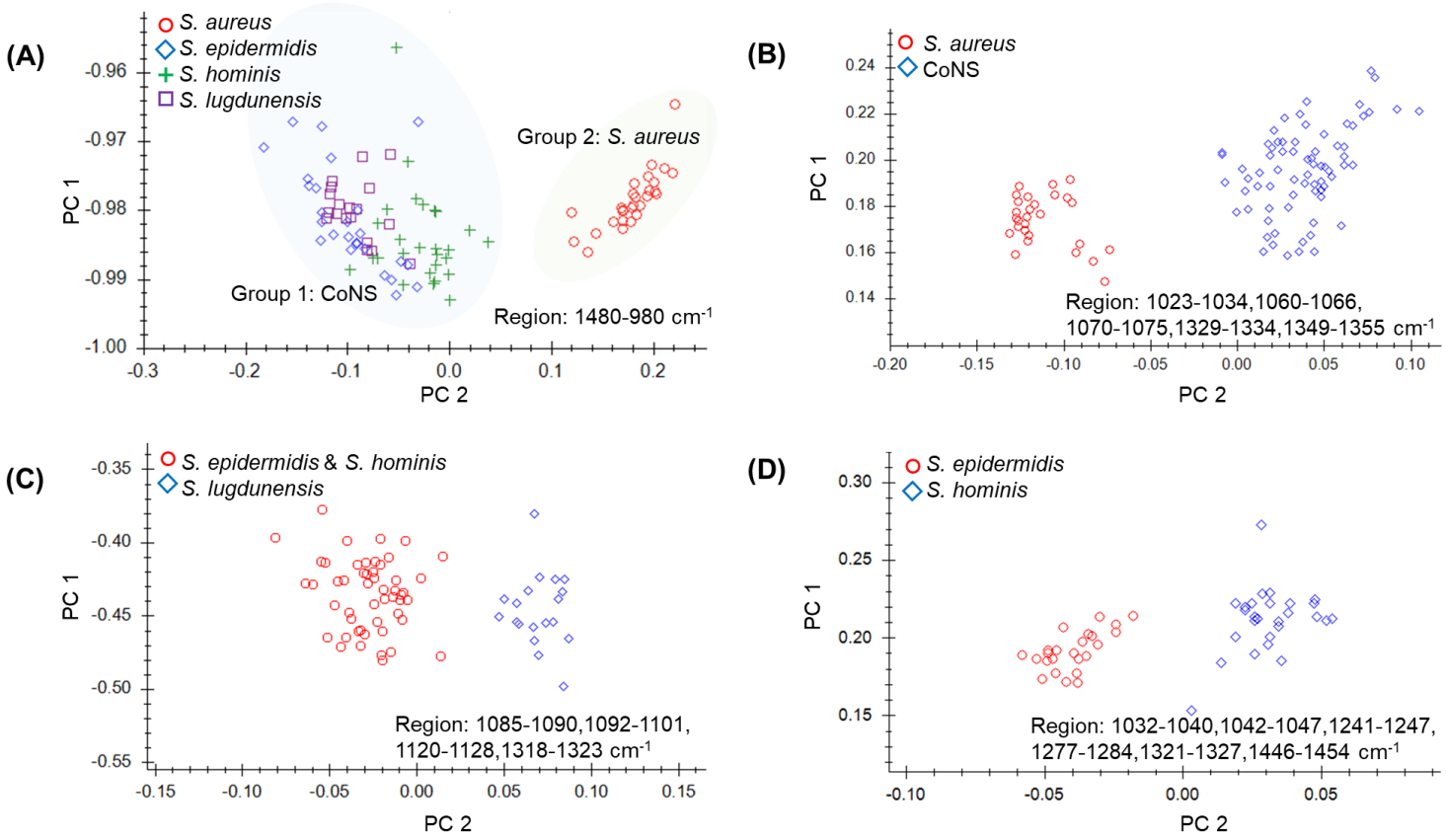


Figure 3.16. Principal component analysis' principal component score plots of ATR-FTIR spectral groups in pairs: (A) broad region observation of 4 *Staphylococcus* species naturally clustering in 2 groups, (B) after region selection, optimal regions were obtained for the discrimination of *S. aureus* and coagulase negative staphylococci (CoNS), (C) discrimination of CoNS species and (D) final discrimination between CoNS species, *S epidermidis* and *S. hominis*

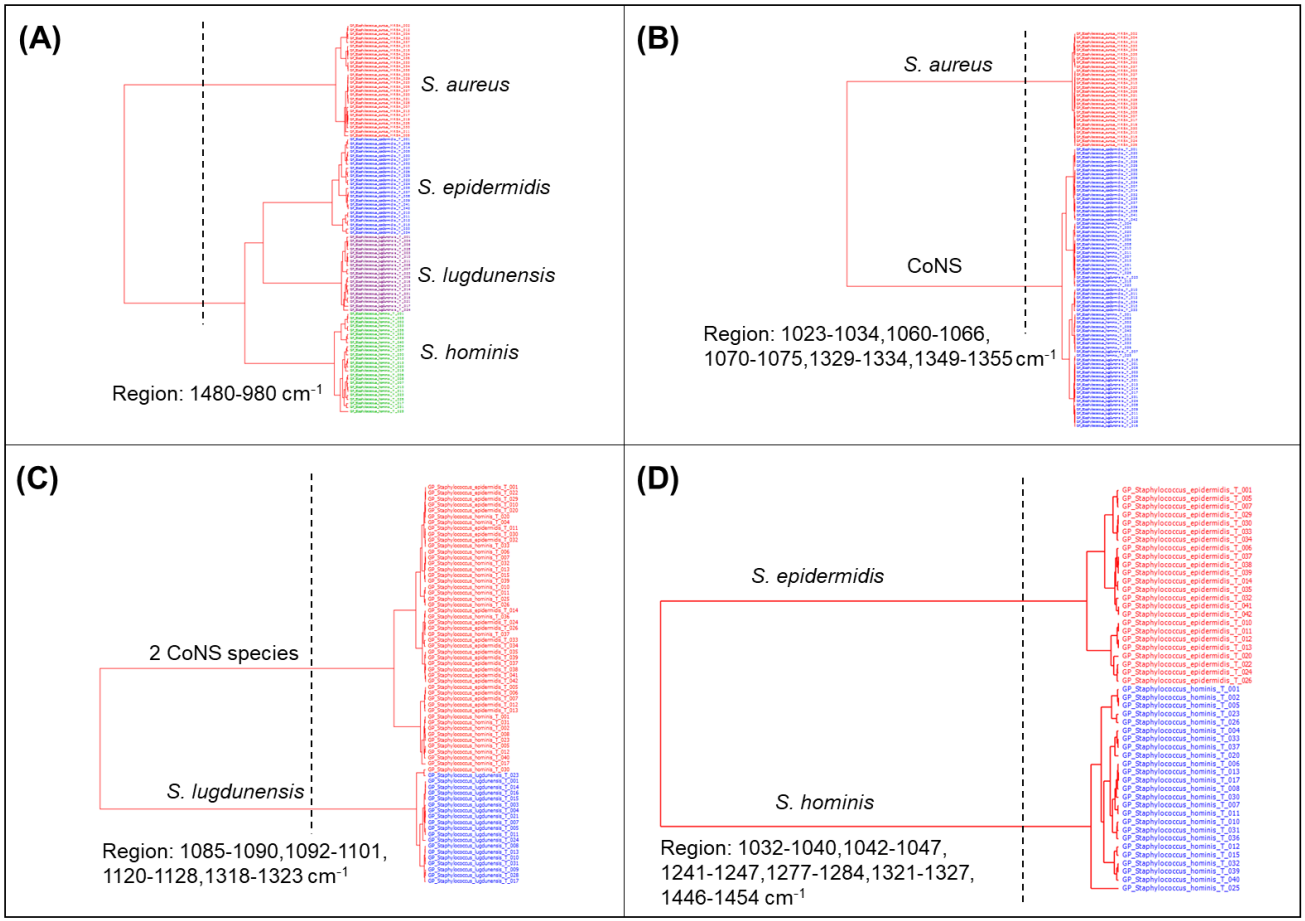


Figure 3.17. Hierarchical cluster analysis dendrogram (cosine similarity distance and ward linkage) of ATR-FTIR spectral groups in pairs: (A) using a broad spectral region (1480-980 cm^{-1}) to differentiate the 4 *Staphylococcus* species into 2 groups, (B) Using optimal spectral regions for the discrimination of *S. aureus* and CoNS, (C) discrimination of CoNS species and (D) final discrimination between CoNS species, *S epidermidis* and *S. hominis*

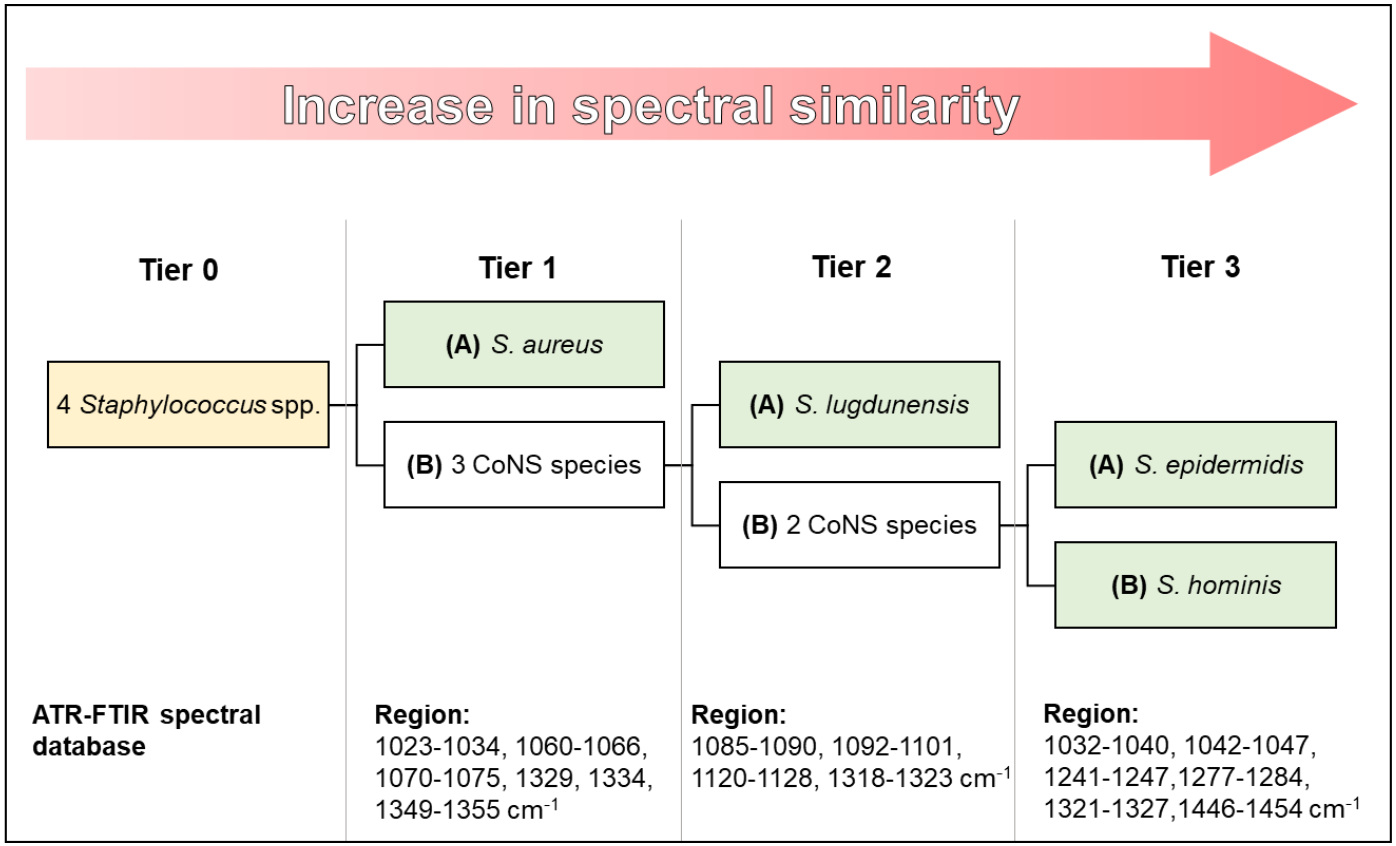


Figure 3.18. Example of sequential multitier pairwise structure for ATR-FTIR spectral database construction for microorganism identification where each pair in each Tier requires a specific spectral region (generated from a feature selection algorithm) for discriminating between each pair into two distinct groups. Pairs of spectra closest to Tier 0 has higher spectral dissimilarity compared to those pairs at a higher Tier level.

3.6.2.2. II: Application of K nearest neighbor (k-NN) algorithm for species identification

Once the spectral database structure is determined with the use of optimized spectral regions, the decision-like tree is implemented using the spectral pairs and their respective spectral regions using an in-house written software; MultiLevel Classifier (MLC). The MLC software holds the spectral database which is constructed to split into multiple folders representative of each tier as shown in Figure 3.18. The software executes a stepwise separation, where a spectrum of an unknown isolate is used as an external input and is assigned to one of the groups in Tier 1. Subsequently, the assigned group from Tier 1 is used as an input for Tier 2 group separation and so on until the spectrum of the unknown isolate is assigned to a single group (e.g., Gram-stain type, genus, species, serotype or serovar).

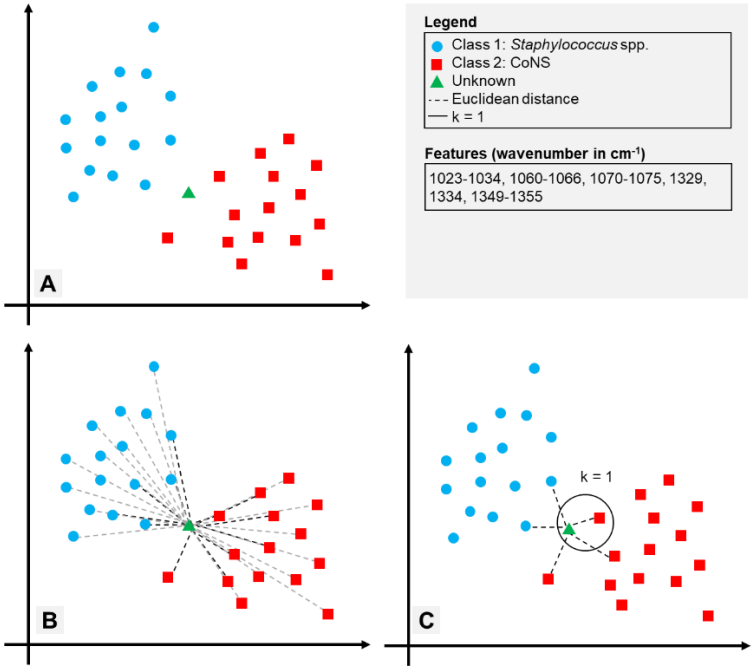


Figure 3.19. Example of classification of an unknown spectrum between two predefined sets of spectra (Class 1 and Class 2). (A) k-NN plot of spectral reference database of Class 1 and Class 2, and the unknown spectrum based on the predetermined features. Subsequently, (B) is the visual representation of the Euclidean distance from the unknown point to the different spectra in each class. Lastly, (C) represents the k-NN when k=1, as such the unknown has the shortest distance to one spectrum in Class 2; classifying the unknown as CoNS

In summary and using the tier wise structure in Figure 3.18, the first pair, *S. aureus* and CoNS are discriminated using the spectral region 1023-1034, 1060-1066, 1070-1075, 1329, 1334, 1349-1355 cm^{-1}, and employing the MLC software, predictions are achieved by loading the spectral reference database and selecting a spectrum or spectral directory of unknowns for prediction, within seconds a spreadsheet is exported with the top three closest match, along with the spectral similarity distance to the matched spectrum and the SD from the reference mean of the predicted (Figure 3.20).

	A	B	C	D	E	F	G	H	I	J	K	L
1	Date: 2020-10-25 07:09											
2	Analysis folder: C:\Users\Administrator\Desktop\20200725 - multicenter	Hit 1			Hit 2			Hit 3				
3	Unknown	Genus	Species	Distance	Genus	Species	Distance	Genus	Species	Distance	SDs From Mean	Confidence
4	YT_Candida_albicans_T_SAB_AE_CHUM_259_ATRC3_20200924_L1	Candida	albicans	0.6848	Candida	albicans	0.673	Candida	albicans	0.6072	3.796404201	Inconclusive
5	YT_Candida_albicans_T_SAB_AE_CHUM_259_ATRC3_20200924_L2	Candida	albicans	0.6435	Candida	albicans	0.6394	Candida	albicans	0.6304	0.197603933	Medium
6	YT_Candida_albicans_T_SAB_AE_CHUM_259_ATRC3_20200924_L3	Candida	albicans	0.7544	Candida	albicans	0.6921	Candida	albicans	0.6551	0.093413445	Medium
7	YT_Candida_dubliniensis_T_SAB_AE_HMR_950_ATRC3_20200925_L1	Candida	dubliniensis	0.9307	Candida	dubliniensis	0.9149	Candida	dubliniensis	0.8707	1.740677555	Medium
8	YT_Candida_dubliniensis_T_SAB_AE_HMR_950_ATRC3_20200925_L2	Candida	dubliniensis	0.9529	Candida	dubliniensis	0.9353	Candida	dubliniensis	0.9328	-0.305371527	High
9	YT_Candida_dubliniensis_T_SAB_AE_HMR_950_ATRC3_20200925_L3	Candida	dubliniensis	0.946	Candida	dubliniensis	0.9247	Candida	dubliniensis	0.8951	1.708556757	Medium
10	YT_Candida_glabrata_T_SAB_AE_HMR_930_ATRC3_20200925_L1	Candida	glabrata	0.7661	Candida	glabrata	0.6732	Candida	glabrata	0.6421	-0.361172501	Medium
11	YT_Candida_glabrata_T_SAB_AE_HMR_930_ATRC3_20200925_L2	Candida	glabrata	0.7342	Candida	glabrata	0.7072	Candida	glabrata	0.682	-0.264602409	Medium
12	YT_Candida_glabrata_T_SAB_AE_HMR_930_ATRC3_20200925_L3	Candida	glabrata	0.7694	Candida	glabrata	0.7027	Candida	glabrata	0.6579	-0.230318013	Medium
13	YT_Candida_krusei_T_SAB_AE_GLEN_033_ATRC3_20201004_L1	Candida	krusei	0.8825	Candida	krusei	0.8215	Candida	krusei	0.809	0.280145259	Medium
14	YT_Candida_krusei_T_SAB_AE_GLEN_033_ATRC3_20201004_L2	Candida	krusei	0.9351	Candida	krusei	0.9214	Candida	krusei	0.9148	0.246209517	High
15	YT_Candida_krusei_T_SAB_AE_GLEN_033_ATRC3_20201004_L3	Candida	krusei	0.9099	Candida	krusei	0.8958	Candida	krusei	0.881	0.512075165	High
16	YT_Candida_tropicalis_T_SAB_AE_GLEN_063_ATRC3_20201004_L1	Candida	tropicalis	0.9185	Candida	tropicalis	0.9111	Candida	tropicalis	0.9077	-0.340561643	High
17	YT_Candida_tropicalis_T_SAB_AE_GLEN_063_ATRC3_20201004_L2	Candida	tropicalis	0.9876	Candida	tropicalis	0.9859	Candida	tropicalis	0.9761	-0.257724525	High
18	YT_Candida_tropicalis_T_SAB_AE_GLEN_063_ATRC3_20201004_L3	Candida	tropicalis	0.9766	Candida	tropicalis	0.9744	Candida	tropicalis	0.9721	-0.651984663	High

Figure 3.20. Screenshot of prediction output using the Multilevel Classifier application illustrating the spectral filenames, top 3 hits, cosine distance similarity from top hit, standard deviation from mean spectrum of top hit and the confidence of the top hit

Each of three triplicate spectra of an unknown sample are analyzed using the multilevel spectral reference database until there are no more tiers to interrogate (terminal folder). Again, in reference to Figure 3.18, the unknown spectrum enters the spectral reference database at Tier 0 and moves onto Tier 1 with the decision of choosing between (A) *S. aureus* or (B) CoNS. Figure 3.19 illustrates the process of determining which group the unknown spectrum is assigned by the k-NN algorithm. First, spectra in Class (A) and Class (B) are plotted using spectral features in the 1023-1034, 1060-1066, 1070-1075, 1329, 1334, 1349-1355 cm^{-1} spectral ranges, second, the unknown spectrum is introduced into the plot where the Euclidean distance is calculated between each point to the unknown spectrum. With a set parameter of k=1, the k-NN is determined by the lowest Euclidean distance. As such, the unknown spectrum has the shortest distance to a spectrum in Class 2, therefore, the unknown spectrum is classified as (B) or CoNS and will proceed into Tier 2 where the same analysis is continued until there are no more tiers to interrogate. If the k-NN, (k=1), has a closest distance to Class 1, then the unknown

spectrum would have been classified as (A) *S. aureus*; no further Tiers are interrogated (terminal folder).

3.6.2.3. Confidence limits parameters for spectrotyping-based predictions

Once a classification/prediction is determined, confidence of the classification is essential to validate whether the results are reliable or are non-reportable. The confidence is determined by 2 factors, the cosine similarity distance, and the SD of the classified to the mean of those spectra in the reference database. In Figure 3.21, the unknown spectrum has a SD <1 and the cosine similarity distance, range between -1 and 1, where values closest to -1 or 1 is the closest distance to one group or the other. Values closest 0 means (orthogonal), provide unreliable assignment to either group. Empirical confidence limits (based on the analysis of ~18,000 spectra acquired from ~6,000 isolates) of SD values >3 and/or cosine distances close to 0 are considered inconclusive.

The spectrum of an unknown microbial sample is compared in a sequential manner (left-right or top-down) between the pairs until it reaches the best match, outputting a report with the spectral distance, SD from the mean and the confidence of the match.

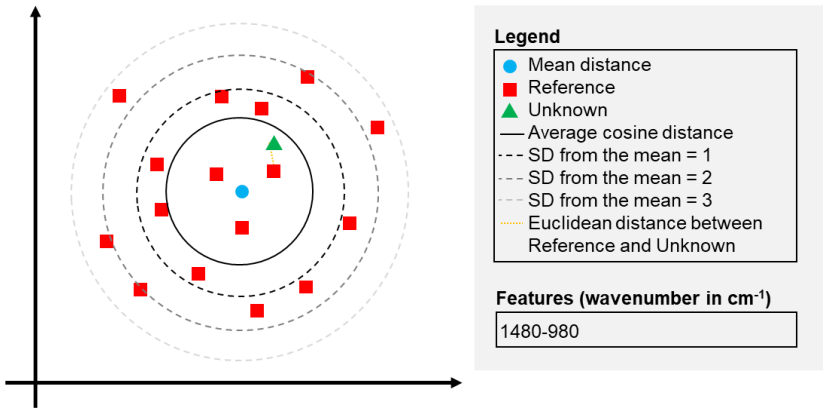


Figure 3.21. Demonstration of the Multilevel Classifier confidence level of classification through the standard deviation (SD) of the unknown spectrum from the mean spectrum of the predicted

Confidences are based on pre-defined acceptance limits of SD from the reference mean spectra of the predicted and spectral similarity distances and are outputted as either inconclusive (not providing a sample identification), high, medium or low confidences. Based on the assigned confidence limits for the species, the output result will be reported

as inconclusive (following rules 2 and 3 in Table 3.4) rather than reporting a wrong species identification (i.e., false identification).

Table 3.4. Example of confidence limit determination for the ATR-FTIR spectroscopy-based technique

Example of Validation and Confidence Determination Table		
Rule	Confidence	Confidence limits
1	Inconclusive	If all predictions are different (of three replicates)
2	Inconclusive	SD > 3
3	Inconclusive	Distance < 0.5
4	High	SD ≤ 0.1.5 and distance ≥ 0.78
5	Medium	SD > 1.5 and ≤ 3 AND distance < 0.8 and ≥ 0.6
6	Low	SD > 3 AND distance < 0.6 and ≥ 0.5

* Rules are applied in sequential order

Any isolate reporting as inconclusive will be further investigated and re-identified by ATR-FTIR spectroscopy and the routine method of microorganism identification, either commercially available *in vitro* MALDI-TOF MS diagnostic system or VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile). If results are discordant after reanalysis on both methods, it is necessary to identify the microorganism based on a third method (i.e., gene sequencing) to ascertain the true identification to evaluate the performance of the ATR-FTIR-based method.

3.6.3. Support vector machine as a tool for ATR-FTIR spectral database construction for microorganism identification

Spectrotyping through the feature selection algorithm may not be adequate for certain levels of pairwise discrimination such as between various pairs of Gram-negative bacteria (especially for those species belonging to the *Enterobacteriaceae* family), where they are biochemically similar. Other machine learning algorithms such as artificial neural network may also be explored, but they have the drawbacks of not providing specific spectral regions that may be useful for discovering biomarkers responsible for the discrimination of groups of microorganisms. As such, an SVM algorithm has been employed and evaluated for microorganism identification based on ATR-FTIR spectral data.

SVM algorithms provide accurate results with low computational power and time, however, unlike spectrotyping, they require large datasets to ensure enough

data/representation for both the training and validation sets. As a result of the machine learning algorithm, it is an extremely powerful and elegant tool for rapid classification. The purpose of SVM is to determine the most optimal hyperplane (largest distance/margin between classes) that can successfully segregate 2 or more sets of data. The hyperplane is also known as the decision boundary where the unknown(s) will be classified as either-or classes dependent on which side of the decision boundary it falls into. As such, SVM without further datamining does not provide unclassified responses.

3.6.3.1. Method for SVM prediction model construction

JMP® Pro 15 (SAS Institute Inc, NC) software was employed for SVM model development. After spectral filtration and preprocessing, the spectral data (both database and unknowns (test set)) are imported into JMP® Pro 15 (SAS Institute Inc, NC) as a spectral data matrix. For preparation of the SVM model, the data must be separated into training, validation and test sets where validation and test sets are external from the training set. For example, there are 327 spectra of *C. albicans* and 109 spectra of *C. glabrata* that will be used to train and validate the spectral reference database. External from both the training and validation sets, there are 60 spectra of each species (totalling to 120 spectra) reserved for the testing set (typically, the testing would be the unknown samples collected in routine). Between the training and validation sets, there will be a 75:25 stratified validation split to compensate the imbalance of spectra per species.

Unlike the MLC method which uses a region selection, SVM uses the broad spectral region 1480-980 cm^{-1} as the features, linear kernel function (creates a linear hyperplane to separate the classes) and a cost value of 1 (where cost is defined as a value greater than 0, where increasing the cost value increases the trade-off values between variance and bias which may lead to overfitting the data). The results are reported as the predicted, most likely predicted, in a confusion matrix, misidentification rate of the training, validation and test sets, and the fit details (Figure 3.22).

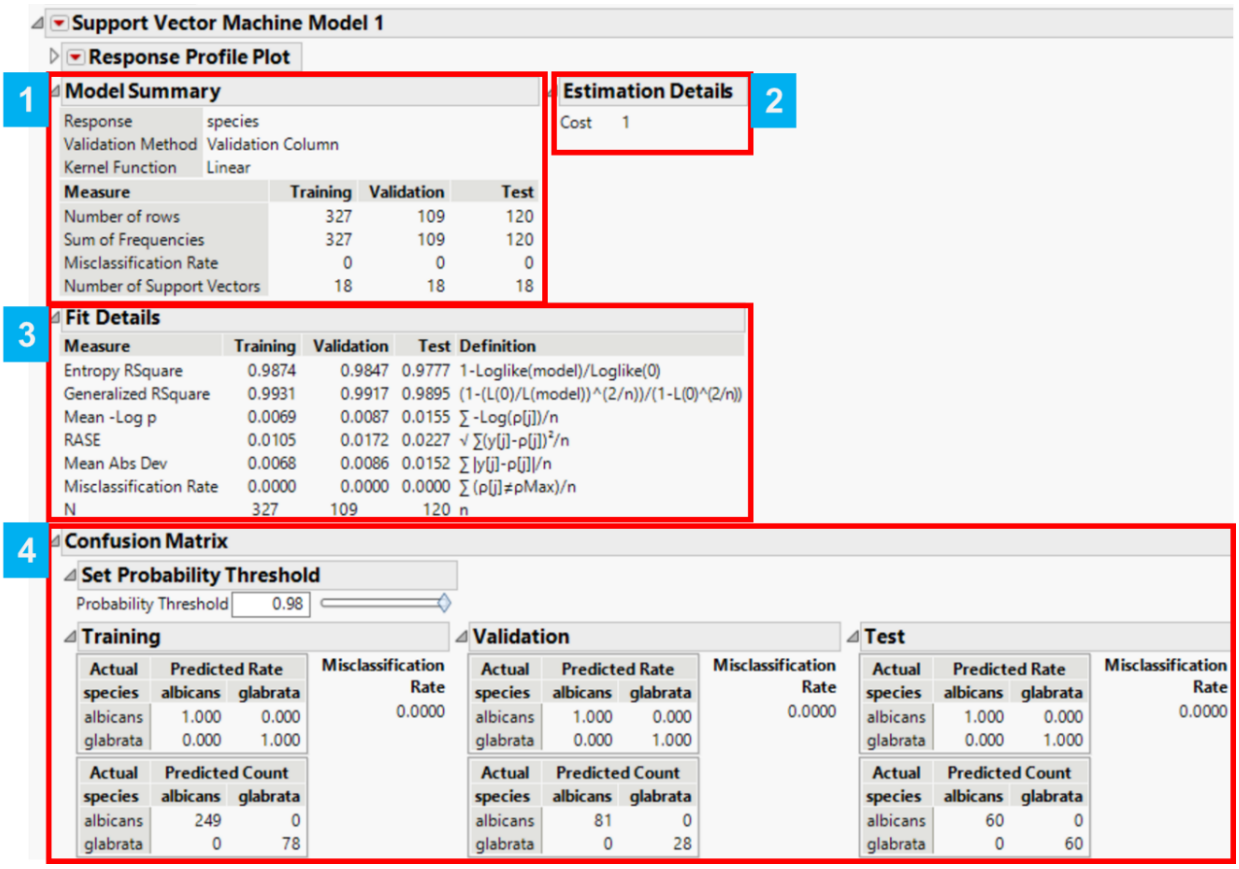


Figure 3.22. Screenshot of JMP® Pro 15 (SAS Institute Inc, NC) support vector machine model output for the classification of *Candida albicans* and *Candida glabrata.* Details for Model Summary (panel 1), Estimation Details (panel 2), Fit Details (panel 3) and Confusion Matrix (panel 4) are further explained in Table 3.5

By default, the misclassification rate is calculated at the probability cut-off of 0.5. In a binary response SVM model, the probability threshold can be manipulated, however, manual recalculation of the misidentification rates is required. The SVM predicted class is determined by where the test spectra are being classified in either-or class, while the likely predicted class is based on a probability cut-off of 0.5. In addition, SVM also allows for multiclass classification, where the SVM predicted class is based on a pairwise combination of the classes and the class that is predicted the most between the pairwise combination of classes is the SVM predicted class.

Table 3.5. Detailed description of numbered panels in Figure 3.22. Screenshot of JMP® Pro 15 (SAS Institute Inc, NC) support vector machine (SVM) model output for the classification of *Candida albicans* and *Candida glabrata*

Panel number	Description
1: Model summary	General summary of the SVM model in terms of displaying information on the selected "Response" (e.g., species which consists of Candida albicans and Candida glabrata), the validation method and the type of kernel function selected. JMP® Pro 15 is based on data within a matrix where each row is a different sample, in this case, a spectrum associated with an isolate. As such, the model summary summarizes how the data was divided into the training, validation and test sets along with the misidentification rates in each set and the number of support vectors generated for classification between the responses.
2: Estimation details	Panel with the details and parameters used in the model such as the cost where cost is defined the value which is associated with margins of the SVM model which attributes to training errors. A larger cost value is associated with narrow margins for a conservative model with few misclassifications while a small cost value is associated with a larger margin, allowing for more misclassification.
3: Fit Model	The Fit Model panel provides the following statistics for the training set, and for the validation and test sets if they are specified (definitions were taken from SAS Institute Inc., Cary, NC online resources): Entropy RSquare: Is a range between 0 to 1, where a value closer to 1 represents the best fit of the model in comparison to the goodness to fit (log-likelihood) of the fitted model to the constant probability model. Generalized RSquare: Measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Values closer to 1 indicate a better fit. Mean -Log *p*: The average of -log(*p*), where *p* is the fitted probability associated with the event that occurred. Smaller values indicate a better fit. RASE (Root Average Square Error): The square root of the mean squared prediction error. Mean Abs Dev: Smaller values indicate a better fit. The average of the absolute values of the differences between the response and the predicted response. The differences are between 1 and *p*, the fitted probability for the response level that actually occurred. Misclassification Rate: The rate for which the response category with the highest fitted probability is not the observed category. N: The number of observations.
4: Confusion Matrix	A confusion matrix is shown for the training, validation, and test of the specified model and represents a two-way classification of actual and predicted responses to evaluate the misclassification rates to evaluate your model.

Although the probability cut-off can be manipulated (increased or decreased), through evaluation, the microbial identification by the ATR-FTIR spectroscopy-based method, would require species specific cut-offs. With that in mind, the spectral reference database may continuously be changing and modified, where the species-specific cut-offs are therefore fluid and inconsistent. However, the probability may be an excellent indicator of confidence levels, whether a prediction is reliable or not; being classified as misidentified or inconclusive.

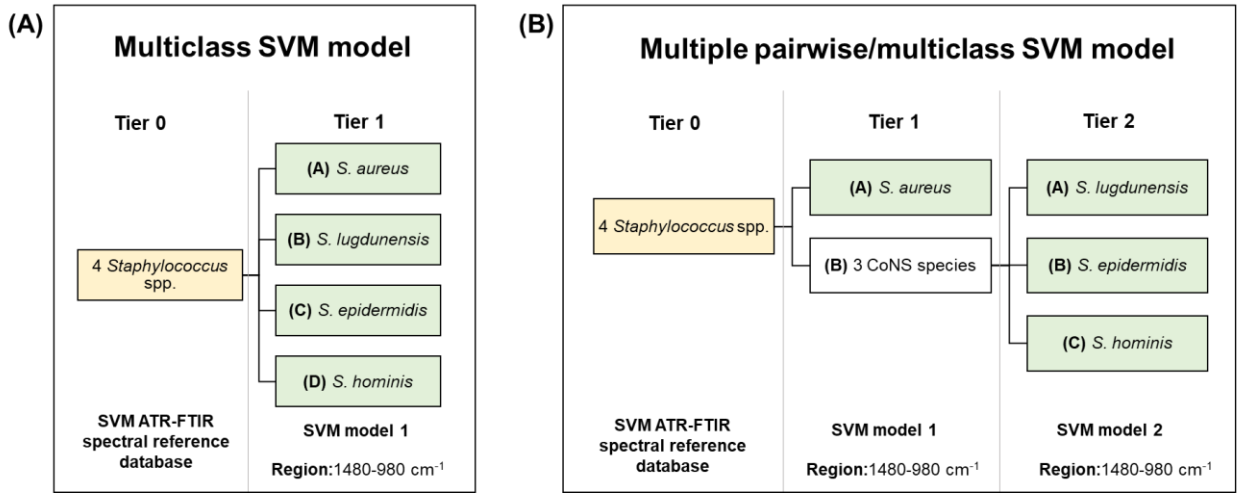


Figure 3.23. Example of SVM model construction based on ATR-FTIR spectra where (A) depicts a single SVM model for the classification of 4 species and (B) depicts a combination of pairwise and multiclass SVM models for the classification of 4 species (the same as in (A))

Similar to the MLC method, multiple SVM models can be stringed together, however, it differs from MLC as it is not limited to pairwise classification, and it is a much more powerful tool compared to MLC in combination with feature selection. The final SVM reference model can therefore be constructed with multiple SVM models in pairwise (binary) or multiclass or a combination of the two (Figure 3.23).

3.6.3.2. Confidence limits parameters for SVM-based predictions

SVM is a model where unknowns are classified into one of the classes defined in the training set. For microorganism identification, there is a need for determining whether a prediction is reliable or not and why is it not reliable – was it misidentified? Or is it inconclusive (not represented in the model, spectra from mixed culture, or the quality of

the spectra are abnormal). The final outputs therefore require confidences levels (or values) to assess the significance of the predictions.

As part of the thesis's work, confidence limits were investigated for the SVM models, where the limits are dependent on a multitude of factors such as: final prediction of the SVM model, the details of the final predictions which includes the predicted class (based on the either-or classification), most likely predicted (based on a probability cutoff of 0.5) and the threshold predicted (based on a self-assigned probability cutoff; variable depending on the model). While the SVM models output predictions as both "Predicted" and "Most Likely", there can be multiple predictions within the Predicted output. For example, a test spectrum can result in Predicted and Most Likely as Class A from a training set that has Class A, Class B and Class C. For the Predicted output, it is based on the prediction results between the combinations: Class A/Class B, Class A/Class C and Class B/Class C, and the Most Likely is based on the highest probability between prob(Class A), prob(Class B) and prob(Class C). As such, for the sake of the example, the test spectrum is predicted as Class A, Class B and Class C in respective order to the pairs in the previous sentence. For the Most Likely predicted class, the SVM model output: prob(Class A)=0.83, prob(Class B)=0.24 and prob(Class C)=0.61. Based on the results, the Predicted and Most Likely predicted is Class A. In an event where there is a prediction of a third class (i.e., Class B) in the multiclass SVM model, the prediction is by default, inconclusive and should not be reported. Spectral investigation of the unknown spectra will be required to ascertain the inconclusive result.

In creating a confidence limit chart consideration of the SVM predicted outputs, predicted probability and threshold predicted probability must be undertaken using 2 different confidence limits charts, (1) for pairwise SVM and (2) for multiclass SVM models.

3.6.3.2.1. Pairwise confidence limits

In both pairwise and multiclass SVM models, the validation set determines the probability threshold of the spectral group (class) and is generally over 0.5. The spectral database developer will have to set probability threshold based on previous experience and discretion to avoid overfitting or underfitting the data. The use of the probabilities of the other groups of spectra may aid in selecting a probability threshold.

Table 3.6 lists 6 genera and 2 species of microorganisms. These 8 spectral groups can be paired as one genus (or species) versus the 7 others (defined as the 2nd group). For example, SVM model 1 is a pair consisting of spectra belonging to *Achromobacter* and the combined remaining groups of the second tier (i.e., *Kingella*, *Salmonella*, *Serratia*, *Shigella*, *Stenotrophomonas*, NFGNB and *Enterobacteriaceae*). The SVM model 2 is then *Kingella* versus the other groups in tier 2, including *Achromobacter*, and so on and so forth. The details of the example and database is further described in CHAPTER 4. The probability thresholds based on the results of the SVM validation are reported in Table 3.6 and are used to determine the prediction confidences.

Table 3.6. Example of support vector machine (SVM) probability threshold for pairwise models between the microorganism/group and the combined spectra of the remaining groups of the spectral database

Microorganism/Group	SVM probability threshold (p-value)
Achromobacter	0.8043
Kingella	0.9022
Salmonella	0.9022
Serratia	0.6739
Shigella	0.8043
Stenotrophomonas	0.8261
NFGNB	0.8043
Enterobacteriaceae	0.5761

The prediction confidences are based on a scoring system ranging from 0 to 100 where the points are based on weighted values of the SVM model outputs (predicted, most likely predicted and threshold predicted). For those predicting the same class for both the Predicted and Most Likely Predicted, a high confidence score is from 87.5 to 100, medium ranges from 70 to 78.4, low ranges from 52.5 to 65 and inconclusive results ranges from 0 to 52. Table 3.7 and Table 3.8 detail the weights provided to the type of output obtained from the SVM model, the global output (Predicted and Most Likely) and the SVM prediction breakdown (Predicted, Most Likely and Threshold Predicted). Fraction predicted simply indicates how many of the same prediction in the SVM prediction breakdown was obtained for the predicted. For example, the SVM model global output (for a pair; 2 outputs; class A and B), results for Predicted and Most Likely predicted as

class A, however, in the detailed SVM prediction breakdown (3 outputs), class A was predicted for Predicted and Most Likely, but was not predicted for the Threshold Predicted; it was only predicted as class A two out of three times (2/3; fraction predicted). Moreover, in continuation of the example and in reference to Table 3.7, the SVM prediction breakdown (row 6) also resulted in B being Predicted and Most Likely Predicted as Class B. As such, a total score of 70 (50 for global A prediction and 20 points total for predicting A and B; score is out of 100) is obtained for the confidence of the unknown being predicted as A (medium level of confidence) and a score of 20 for the unknown being predicted as B (inconclusive prediction). As such, for the given example, the better of the two predictions is A with medium level of confidence.

Moreover, Table 3.8 describes prediction confidences for SVM models with prediction outputs for Predicted and Most Likely predicted are not the same. Unlike those with prediction outputs that are the same, there are no high confidence predictions for those that are not the same. In reference to Table 3.8, medium confidence scores range from 70-75, low confidence level scores from 53-65 and inconclusive results range from scores of 10-50. In both tables, the prediction confidence scores are color coordinated by the level of confidence, red for inconclusive, yellow for low confidence, blue for medium confidence and green for high confidence. Typically, only medium and high confidence scores should be reported while low confidence scores should be revisited as well as inconclusive results.

Table 3.7. Assigned weights for pairwise SVM model for the confidences of microorganism identification by the ATR-FTIR spectroscopy-based method predictions: Global SVM predicted as A and is most likely A

SVM prediction output		SVM prediction breakdown								Prediction Confidence score	
A	A	Predicted A				Predicted B					
Predicted	Most Likely	Predicted	Most Likely	Threshold Predicted	Fraction predicted	Predicted	Most Likely	Threshold Predicted	Fraction predicted	A	B
25	25	8.25	8.25	8.5	3/3	8.25	8.25	8.5	3/3	75	25
25	25	10	10	10	3/3	10	10	-5	2/3	80	15
25	25	12.5	12.5	12.5	3/3	12.5	-5	-5	1/3	87.5	2.5
25	25	16.65	16.65	16.7	3/3	-	-	-	0/3	100	0
25	25	10	10	-5	2/3	10	10	10	3/3	65	30
25	25	12.5	12.5	-5	2/3	12.5	12.5	-5	2/3	70	20
25	25	16.7	16.7	-5	2/3	16.6	-5	-5	1/3	78.4	6.6
25	25	25	25	-5	2/3	-	-	-	0/3	95	0
25	25	12.5	-5	-5	1/3	12.5	12.5	12.5	3/3	52.5	38
25	25	16.6	-5	-5	1/3	16.7	16.7	-5	2/3	56.6	28
25	25	25	-5	-5	1/3	25	-5	-5	1/3	65	15
25	25	50	-5	-5	1/3	-	-	-	0/3	90	0
25	25	-5	-5	-5	0/3	16.65	16.65	16.7	3/3	35	50
25	25	-5	-5	-5	0/3	25	25	-5	2/3	35	45
25	25	-5	-5	-5	0/3	50	-5	-5	1/3	35	40
25	25	-5	-5	-5	0/3	-	-	-	0/3	35	0
25	25	16.65	16.65	16.7	3/3	-	-	-	0/3	100	-
25	25	16.65	16.65	-5	2/3	-	-	-	0/3	78.3	-
25	25	16.65	-5	-5	1/3	-	-	-	0/3	56.7	-
25	25	-5	-5	-5	0/3	-	-	-	0/3	35	-

Table 3.8. Assigned weights for pairwise SVM model for the confidences of microorganism identification by the ATR-FTIR spectroscopy-based method predictions: Global SVM predicted as A and is most likely B

SVM prediction output		SVM prediction breakdown								Prediction Confidence Scores	
A	B	Predicted A				Predicted B					
Predicted	Most Likely	Predicted	Most Likely	Threshold Predicted	Fraction predicted	Predicted	Most Likely	Threshold Predicted	Fraction predicted	A	B
25	25	8.25	8.25	8.5	3/3	8.25	8.25	8.5	3/3	50	50
25	25	10	10	10	3/3	10	10	-5	2/3	55	40
25	25	12.5	12.5	12.5	3/3	12.5	-5	-5	1/3	62.5	28
25	25	16.65	16.65	16.7	3/3	-5	-5	-5	0/3	75	10
25	25	10	10	-5	2/3	10	10	10	3/3	40	55
25	25	12.5	12.5	-5	2/3	12.5	12.5	-5	2/3	45	45
25	25	16.7	16.7	-5	2/3	16.6	-5	-5	1/3	53.4	32
25	25	25	25	-5	2/3	-5	-5	-5	0/3	70	10
25	25	12.5	-5	-5	1/3	12.5	12.5	12.5	3/3	27.5	63
25	25	16.6	-5	-5	1/3	16.7	16.7	-5	2/3	31.6	53
25	25	25	-5	-5	1/3	25	-5	-5	1/3	40	40
25	25	50	-5	-5	1/3	-5	-5	-5	0/3	65	10
25	25	-5	-5	-5	0/3	16.65	16.65	16.7	3/3	10	75
25	25	-5	-5	-5	0/3	25	25	-5	2/3	10	70
25	25	-5	-5	-5	0/3	50	-5	-5	1/3	10	65
25	25	-5	-5	-5	0/3	-5	-5	-5	0/3	10	10

3.6.3.2.2. Multiclass confidence limits

Unlike pairwise confidence levels, multiclass confidence limits are based solely on the SVM model outputs (Predicted and Most Likely Predicted) and the threshold probability. Similar to pairwise SVM models, per class, a probability threshold based on the validation set of the multiclass SVM model is determined. The Most Likely Predicted is the class with the highest probability from all class's prediction probabilities. As such, to determine the confidence of the prediction, the Predicted class, the Most Likely predicted class and the threshold probability are examined.

Table 3.9. Assigned confidence levels for multiclass SVM models for microorganism identification by the ATR-FTIR spectroscopy-based method predictions where global SVM predicted as A and varying most likely predicted

Predicted	Most Likely Predicted	Confidence level
A	A and p > probability threshold	High
A	A and p < probability threshold	Medium
A	B and p > probability threshold	Low
A	B and p < probability threshold	Inconclusive

Based on Table 3.9, if the Predicted class is A (from multiple classes, A, B, C, D, E…etc.) and the Most Likely Predicted is A and the prob(Class A) is greater than the probability threshold, than the confidence is high; if the Most Likely Predicted is A and the prob(Class A) is less than the probability threshold, than the confidence of the prediction is medium. Moreover, still with the Predicted as class A, if the Most Likely Predicted is B (or any other class) with a prob(Class B) greater than the probability threshold, than the confidence level is low. If the Most Likely Predicted is B with a prob(Class B) less than the probability threshold, than the result is non-reportable – inconclusive.

3.7. CONCLUSION

The current chapter demonstrates the process of analysis of clinical microbial samples-based ATR-FTIR spectroscopy for microorganism identification. The outlined method is easily implementable in clinical routine and uses already used isolation and culture media found in the clinical microbiology laboratory. Multiple manufacturers of ATR-FTIR spectrometers are available and through the studies presented in the chapter,

instrument-to-instrument variability from the same model are insignificant when creating a spectral database for microorganism identification. Though instrument-to-instrument variability of different models from the same of different manufacturers are more significant than from instruments of the same model, again, the outlined method for spectral database construction eliminates most of the variability. The proposed method also standardizes the operating procedure for reproducible high-quality spectra of microorganisms and outlines available methods for spectral preprocessing and creation of spectral databases using selected algorithms to identify unknown microorganisms. The combined use of a novel spectrotyping and machine learning results in a technique with high computation power, strain identification and typing capabilities. As such, the proposed standardized operating protocol is practical, simple, cost-effective, transferable between laboratories and has great potential to provide microbiology laboratories with a new *in vitro* diagnostic microbial tool.

3.8. REFERENCES

1. Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P-E, Rolain JM, Raoult D. 2009. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clinical Infectious Diseases 49:543-551.

2. Bailey D, Diamandis EP, Greub G, Poutanen SM, Christensen JJ, Kostrzew M. 2013. Use of MALDI-TOF for Diagnosis of Microbial Infections. Clinical Chemistry 59:1435-1441.

3. CLSI. 2008. Abbreviated Identification of Bacteria and Yeast, 2nd Edition doi:1-56238-681-6, p 64.

4. Baldauf NA, Rodriguez-Romo LA, Männig A, Yousef AE, Rodriguez-Saona LE. 2007. Effect of selective growth media on the differentiation of Salmonella enterica serovars by Fourier-transform mid-infrared spectroscopy. Journal of microbiological methods 68:106-114.

5. Kim S, Reuhs B, Mauer L. 2005. Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. Journal of applied microbiology 99:411-417.

6. Kümmerle M, Scherer S, Seiler H. 1998. Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy. Appl Environ Microbiol 64:2207-2214.

7. Maquelin K, Kirschner C, Choo-Smith L-P, van den Braak N, Endtz HP, Naumann D, Puppels G. 2002. Identification of medically relevant microorganisms by vibrational spectroscopy. Journal of microbiological methods 51:255-271.

8. Ghetler A. 2010. Development of an expert system for the identification of bacteria by focal plane array Fourier transform infrared spectroscopyMcGill University.

9. Kirkwood JP. 2007. Identification of Bacteria by Infrared Imaging with the Use of Focal Plane Array Fourier Transform Infrared SpectroscopyMcGill University Libraries.

Connecting statement 2

The outline proposed in the previous chapter requires a standardized method and protocol for spectrotyping and datamining of ATR-FTIR spectra (from clinical sample with confidence of identification/classification) for various applications such as bacterial and yeasts species identification, strain typing and classification of various microorganism complexes.

The standardized protocol is implemented in the following chapter to evaluate the applicability of ATR-FTIR spectroscopy, spectrotyping and SVM as a tool for microorganism identification.

CHAPTER 4. DEVELOPMENT AND EVALUATION OF SPECTROTYPING METHOD FOR MICROBIAL IDENTIFICATION AND TYPING

4.1. ABSTRACT

Fourier transform infrared (FTIR) spectroscopy is a well-established analytical technology that has been used in various fields of study, chemistry, forensics, ecology, medicine, astronomy and more. The mid-infrared region (4000-400 cm^{-1}) is extremely useful for quantifying biochemical composition and identifying chemical bonds by their molecular vibrational bond energy absorption to generate unique spectroscopic "fingerprints". Infrared (IR) spectral acquisition can be simplified by using an attenuated total reflectance (ATR) accessory. For microbiology, an IR spectrum of intact bacterial cells taken from a pure colony from a culture plate, represents the biochemical constituents of the microbial cells (i.e., proteins, lipids, polysaccharides, DNA and RNA, etc.), it has been termed a "whole-organism fingerprint" technique. Given the biochemical differences between different microorganism, ATR-FTIR spectroscopy is capable of differentiation and identification upon creating spectral databases. Clinical isolates were collected over a 3-year period from 9 centers related to clinical microbiology and food surveillance microbiology laboratories in Canada, a research center in Australia and a hospital in the United Kingdom. A total of 7344 isolates were collected where at least triplicate ATR-FTIR spectra of each isolate were acquired using a defined culturing and spectral acquisition method directly from cultured plates. Spectrotyping, the analysis of relative IR absorbance intensities between groups of spectra was achieved for selecting pairs of dissimilar spectral groups in order to construct a multitier pairwise spectral library search database for microorganism identification at the genus and/or species level. Where spectrotyping strained (for those spectra collected from the *Enterobacteriaceae* family), powerful machine learning algorithms such as support vector machines were employed to complete the construction of the search database. The validation of the spectral database's performance resulted in an overall 94.3% and 95.9% correct genus and species identification for those microorganisms represented in the spectral reference database. For the prospective 3-months evaluation in a routine clinical setting, the ATR-FTIR spectroscopy method and microorganism identification technique resulted in 97.2% and 98.4% correct genus and species identification in respective order. As such, the

proposed technique has many benefits such as being compact, easy-to-use, reagent-free, rapid and may be a useful tool for rapid screening to assist current methods.

4.2. INTRODUCTION

Rapid identification of bacteria in hospitals are vital for improving patient outcomes and to improve the efficiency of the current health care system. Nosocomial infections by *Clostridium difficile* and methicillin-resistant *Staphylococcus aureus* (MRSA) may lead to complications and deaths (1). Moreover, non-fermenting Gram-negative bacilli (NFGNB) are a group of GN bacteria that are unable to catabolize glucose and therefore have the inability to ferment sugars. Currently, *Acinetobacter*, *Bordetella*, *Burkholderia*, *Legionella*, *Moraxella*, *Pseudomonas* and *Stenotrophomonas* are classified as NFGNB and account for 15% of all microorganisms identified in clinical routine (2). These microorganisms are challenging to identify and are intrinsically resistant to antibiotics (3). NFGNB lacks distinct phenotypic characteristic for differentiating from one species to another. Unlike the others, *Pseudomonas aeruginosa* are the most phenotypically dissimilar ones from the other species. Identification of NFGNB by conventional biochemical techniques include, colony morphology, glucose fermentation, carbon source, gelatin hydrolysis and others and may be inconclusive, unreliable, time-consuming and require up to several days for results (4). In addition, the biochemical activity of the NFGNB are low where commercial identification methods are reported to produce misidentifications (5).

Cystic fibrosis (CF) is an inherited autosomal recessive chronic genetic disorder of the secretory glands with over 70,000 affected worldwide (6). CF is caused by a mutation in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene which is inherited by both faulty *CFTR* gene carriers' parents. As a result of the gene mutation, CF patients suffers from insufficient movement of water in and out of the epithelial cells. With the poor movement of water between cells, CF patients suffer have very thick and sticky mucus which inevitably obstruct airways and glands. Individuals affected by the disease have a low life expectancy of approximately 37 years with a mortality rate of approximately 80% due to obstructive lung disease and infections (7). Over the recent decades, there has been an emergence of opportunistic NFGNB to cause serious infections in CF patients and with the rise in antibiotic resistance of the microorganisms of these NFGNB -

treatment and diagnosis is also difficult (3). A single drug is incapable of treating all infections and therefore necessitating the need for appropriate, rapid and reliable microorganisms identification (4). Conventionally, *P. aeruginosa* and *Stenotrophomonas maltophilia* are accurately identified by traditional methods, however, species withing the genus *Burkholderia* proved to be more difficult for identification (5).

Furthermore, clinical microbiology laboratories currently identify bacteria via carbon assimilation system such as VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile) and traditional manual biochemical methods. In some circumstances, these phenotypic techniques are unable to provide a reliable identification where genotypic methods such as polymerase chain reaction (PCR)-based methods become necessary. The latter techniques are currently carried out at provincial or national microbiology reference laboratories level; however, genotypic methods are time consuming, require skilled technicians and are costly for routine analysis. More recently, matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) has been proven to be more cost-efficient and rapid than the conventional phenotypic methods (8-10). The MALDI-TOF MS mass spectrum is compared to other spectra in the commercial MALDI-TOF MS spectral database to identify the unknown microbial sample. Although MALDI-TOF MS has recognizable benefits, it requires harsh reagents, is unable to differentiate *Shigella* species from *Escherichia coli*, requires skilled technicians and is a costly investment for small hospitals, clinics and research centers.

Fourier transform infrared (FTIR) spectroscopy is another spectroscopic technique that may also provide the means of microbial identification. FTIR spectroscopy is a well-established analytical technology that has been used in various fields of study: chemistry, forensics, ecology, medicine, astronomy, and more. The mid-infrared region (4000-400 cm^{-1}) is extremely useful for quantifying biochemical composition and identifying chemical bonds by their molecular vibrational bond energy absorption to generate unique spectroscopic “fingerprints”. Infrared (IR) spectral acquisition can be simplified by using an attenuated total reflectance (ATR) accessory. For ATR-FTIR spectroscopy, the IR beam from the IR source is launched at a defined critical angle into an IR transparent crystal (e.g., ZnSe, Ge, Si or a diamond) where total internal reflectance occurs within the

crystal forming an evanescent wave perpendicular to the transmitting IR beam above the sampling crystal. With the sample on the sampling surface, partial attenuation of the evanescent wave by the sample takes place and travels to the detector. An IR spectrum is achieved by taking the ratio between the spectrum of the sample and the spectrum of the crystal in the absence of the sample. For microbiology, an IR spectrum of an intact bacterial cells taken from a pure colony represents the biochemical constituents of the cells (i.e., proteins, lipids, polysaccharides, DNA and RNA, etc.), it has been termed a "whole-organism fingerprint" (11, 12). Given the biochemical differences between different microorganisms, FTIR spectroscopy is capable of differentiation and identification upon creating spectral databases. When FTIR spectroscopy is coupled with the ATR mode of spectral acquisition, FTIR spectroscopy is a low-cost, reagent-free technique that provides results within minutes after initial growth of bacteria in culture.

Not all clinical laboratories have the financial capacity to invest in robust technologies such as MALDI-TOF MS for microbial identification. For example, to-date, Canada's largest mother and child center, the Centre hospitalier universitaire Sainte-Justine (CHUSJ) which specializes in children diagnosed with CF, is unfortunate to not own a MALDI-TOF MS system. Those microorganisms associated with the genetic disease such as those belonging to the *Burkholderia cepacia* complex are difficult to identify and require outsourcing to identify the microorganism to the species level – taking up to weeks to receive results; in turn, delaying appropriate treatment for these patients. As such, the aim of this chapter is to examine the spectral features (spectrotyping) of commonly isolated and difficult to identify microorganisms (e.g., *E. coli* and *Shigella*; *Enterobacteriaceae* family; and species of the *Burkholderia cepacia* complex) to determine spectral similarities between genus, species and complexes.

To the best of the author's knowledge, at the current time, this is the first ever evaluation of ATR-FTIR spectroscopy as a tool for a large spectral set of clinically and food relevant microorganisms and it is the first to prospectively evaluate the developed ATR-FTIR spectroscopy-based microorganism technique in a clinical microbiology laboratory over a 3-month period. Results of this chapter will outline the applicability for a cost-effective alternative to current methodologies without compromising the reliability of the results.

4.3. MATERIALS AND METHODS

4.3.1. Sample collection

All samples for the spectral reference database construction were collected from a total of 9 microbiology laboratories; 5 clinical, 2 clinical reference and 2 food microbiology laboratories stored in 10% glycerol vials stored at -80°C. A total of 7344 isolates were isolated from a wide variety of sources, such as, blood, pus, sputum, urine, stool, wounds, nasal swabs, skin swabs, animal feed, chicken, fruits, and vegetables. Of the clinical isolates (bacteria and yeasts), the isolates were isolated from neonates to seniors, healthy and/or living with a medical condition and/or suffering from current infections.

For the prospective set of spectral data, a total of 391 bacterial isolates were collected directly from routine that were cultured on OXOID Columbia Blood Agar with 5% Sheep Blood (Thermo Fisher Scientific, Nepean, ON) (BAP) over a consecutive 3-month period at the Centre hospitalier universitaire Sainte-Justine hospital in Montreal, Quebec.

Detailed collection of isolates collected for the following chapter are described in Table 4.1, and Table A.6.

4.3.2. Sample preparation

In reference to CHAPTER 3 section 3.3.2, all frozen samples used for the evaluation study were cultured and subcultured on BAP and incubated for 18-24 h at 38 ± 2°C. Isolates are cultured using the 4-quadrant streak plate method for colony isolation. While prospective isolates were cultured on BAP in routine from clinical specimen and incubated with the same conditions as those cultured from frozen.

4.3.3. ATR-FTIR spectral acquisition

Following the protocol outlined in CHAPTER 3 and Table A.5, all spectra (otherwise stated) were collected on a Cary 630 FTIR (Agilent Technologies, CA) spectrometer through direct colony picking with a 1 µL loop and deposition onto the diamond ATR sampling surface. Per sample plate, at least triplicate spectra (different colonies) were acquired using the following spectral acquisition parameters: 64 co-added background scans per spectrum, 64 co-added sample scans, spectral region of 4000-650 cm^{-1}, and

Happ-Genzel apodization with 2 levels of zero filling. After each spectrum that was collected, a lint-free tissue was moistened with 70% ethanol and used to wipe the ATR sampling surface.

Table 4.1. Summary of collected isolates from clinical microbiology laboratories and clinical and food reference microbiology laboratory

Institution	Gram-stain type	No. genera	No. species	No. isolates collected
Clinical microbiology laboratories				
McGill University Heath Center (MUHC)	GN	26	53	830
	GP	15	66	1151
	YT	1	8	349
Centre hospitalier Sainte Justine (CHUSJ)	GN	30	47	902
	GP	11	20	679
	YT	1	3	86
Maidstone Hospital (MAID)[1]	GN	3	3	25
	GP	3	5	40
	YT	1	3	20
Centre hospitalier Sherbrooke (CHUS)	GN	9	7	31
	GP	3	3	10
	YT	1	9	93
Queensland Institute of Medical Research (QIMR)	GN	4	4	12
	GP	7	14	96
TOTAL				4324
Clinical Microbiology Reference Laboratory				
Laboratoire de Santé Publique du Québec (LSPQ)	GN	3	3	5
	GP	2	10	181
	YT	3	15	454
National Microbiology Laboratory (NML)	GP	2	3	263
TOTAL				903
Food Microbiology Reference Laboratory				
Health Canada (HC)	GN	3	4	108
	GP	2	7	24
Canadian Food Inspection Agency (CFIA)	GN	7	6	1812
	GP	4	15	173
TOTAL				2117
TOTAL No. collected isolates				7344

[1]Isolates collected at MAID were acquired on SummitPro Everest™ (Thermofisher Scientific, WI) while the spectra collected on the other institutions were collected on a Cary 630 FTIR (Agilent Technologies, CA)

Spectral quality check, filtration and preprocessing follows those steps outlined in CHAPTER 3 section 3.6.1. In brief, spectra with low biomass, displaying moisture loss (dried samples) and/or exhibiting spectral anomalies were filtered out and each spectrum was preprocessed by vector normalization and the 1st derivative was calculated to increase resolution prior to data analysis for the construction of the spectral database construction. Additionally, after spectral filtration and preprocessing, triplicate spectra per sample analyzed were averaged, resulting in 1 spectrum per sample (or also termed as isolate).

4.3.4. Spectral database construction

The spectral reference database for genus and species identification were constructed using an inhouse software DataAnalysis (McGill University, Quebec) and JMP® Pro 15 (SAS Institute Inc, NC). Spectrotyping is a technique used to determine significant spectral data responsible for differences between two or more groups. In the study, spectrotyping is accomplished with a feature selection algorithm which uses a grid-greed search to determine spectral features responsible (region selection) for discriminating between pairs of similar groups of ATR-FTIR spectra obtained from microorganisms. Pairs of groups of ATR-FTIR spectra are determined with the region selection and the pairs are then stringed together to create a decision-like tree multilevel (or multitier) sequential spectral reference database. Details were previously described in CHAPTER 3 section 3.6.2.

In addition to the spectrotyping method, support vector machine (SVM) algorithm is also employed for genus and species identification where spectrotyping proved to be difficult. SVM algorithms are employed to ascertain that the spectral information is available for datamining using alternative powerful computational tools. Unlike the spectrotyping technique, SVM has a generalization error (or out-of-sample error) which is defined as how accurately the SVM model can predict outcomes for unseen data. Thus, spectral regions are not determined associated with the discrimination and therefore is not considered a spectrotyping method for detecting biochemical differences between spectral groups.

4.3.5. Training, validation, and test sets

Of the total spectra collected from 7344 isolates collected (Table 4.1), 6342 isolates represent bacteria and only 3288 isolates were cultured on identical culture media agar, incubated using the same conditions and were collected on spectrometers of the same make and model. With the spectra collected from 3288 isolates, the data were split into a training set (also referred to as the spectral reference database), evaluation set (a set used to validate the training set) and a test set (independent from the development and validation of the spectral reference database). The evaluation set is a subset of the spectra represented of the training set but is external from the training set. The test set is also external from both the training and evaluation set, however, spectra collected for the test set represents isolates collected in clinical routine over a 3-month period.

Spectra collected from a total of 2619 isolates were used to construct/train and evaluate the spectral reference database, which represents 20 genera of bacteria (13 Gram-negative and 7 Gram-positive bacteria). Of the 20 genera, there are 86 unique species, however, only 27 species are considered represented (at least 14 unique isolates of the species were acquired). Fourteen is the representation cut-off and thus, any species with less than 15 isolates are omitted from obtaining species identification for the evaluation of ATR-FTIR spectroscopy as a tool for microorganism identification. Of the 27 species that are represented, 14 belongs to Gram-negative bacteria and the remaining 13 are Gram-positive bacteria.

In addition to the spectra of bacteria, fungal spectra were also acquired and employed in the initial database construction. Further details of fungal identification by the ATR-FTIR spectroscopy method will be described in CHAPTER 5. For the current chapter, a random group of spectra of yeasts (spectra belonging to 1000 isolates with a wide variety of genera and species) and mold (only 19 isolates were collected for spectral acquisition) were set aside as a set for screening between bacteria and fungi during the evaluation.

From the 3-month prospective collection of spectra of routine isolates, a total of 391 were collected belonging to 31 species where 12 are not represented in the training set.

4.4. RESULTS AND DISCUSSION

4.4.1. ATR-FTIR spectral database construction

Through spectrotyping by region selection, various pairs of spectral groups of microorganisms were selected for the pairwise multitier spectral reference database construction. As such, all spectra from the reserved 2619 isolates were discriminated from fungi spectra using a broad spectral region of 1480-980 cm^{-1}. With the region selection, 1116-1122,1286-1292,1346-1357,1372-1383,1441-1446 cm^{-1} were selected for an optimized discrimination between the two groups. These regions are associated with the C-O stretch of the C-O-C glycosidic linkage and carbohydrates of microorganisms, the amide III component of proteins and, CH_2 and CH_3 bending associated with lipids and proteins (Table 2.2). Continuing the search for the next pair of spectral groups of microorganisms, the first pair of spectral groups were determined, then the second, the third, forth and so on. Upon completion of selecting pairs of spectral groups of microorganisms, the pairs were stringed together to create the spectral reference database structure, but for simplicity, it was separated into three different sets, now referred to as Set A, Set B and Set C (Figure 4.15).

4.4.1.1. Set A: Construction of ATR-FTIR multitier pairwise spectral database for the discrimination of Gram-stain variable, positive and negative bacteria

Set A consists of 6 pairs distributed into 4 tiers, where increased in tier levels are associated with an increase in spectral similarity (Table 4.2 and Figure 4.15). For example, tier 1 is for the discrimination between the domain of bacteria and fungi, which can also be classified on a broader term as the difference between prokaryotic and eukaryotic organisms.

The difference between the two domains narrows down to the biochemical compositional differences such as the cell membrane structure, RNA, DNA, proteins, lipids and carbohydrate composition. The major biochemical differences between the two domains are also observed in the ATR-FTIR spectra in regions associated with the listed biomolecules. While tier 1 demonstrates the most dissimilar groups of spectra, tier 4 of Set A exhibits the most spectrally similar groups of spectra. The final pair of groups of spectra in Set A belongs to *Corynebacterium* spp. and *Micrococcus* spp., which are

Gram-stain variable microorganisms, however, they are classified as Gram-positive bacteria due to the presence of a peptidoglycan layer.

Table 4.2. Set A: Selected regions through a feature selection algorithm for the discrimination between pairs of groups of spectra for the construction of an ATR-FTIR spectral reference database for identifying fungi and Gram-stain variable, Gram-negative and Gram-positive bacteria

Tier	Group	Pair	Spectral region for discrimination (cm^{-1})
1	a	Bacteria	1116-1122,1286-1292,1346-1357,1372-1383,1441-1446
	b	Fungi	
2	c	Yeast	1239-1245,1299-1306,1338-1344,1346-1351,1388-1394,1400-1405,1415-1422,1446-1454
	d	Filamentous fungi[1]	
2	a	Gram-negative bacteria	1090-1096,1114-1120,1137-1142,1230-1267,1323-1342,1353-1360,1409-1415,1450-1456
	b	Gram-stain variable bacteria[2]	
3	a	Gram-positive bacteria	1221-1226,1286-1292,1299-1310,1355-1360,1370-1375,1415-1420,1429-1437,1476-1480
	b	Gram-negative bacteria	
3	c	*Bacillus* spp.	990-995,1116-1122,1154-1159,1185-1191,1258-1267,1278-1284,1286-1292,1295-1305,1347-1355,1385-1390,1407-1415,1469-1474
	d	*Corynebacterium* spp. & *Micrococcus* spp.	
4	a	*Corynebacterium* spp.	990-1005,1057-1088,1137-1169,1180-1196,1230-1383,1385-1401,1403-1416,1437-1457,1459-1470
	b	*Micrococcus* spp.	

[1] Molds
[2] *Bacillus* spp., *Corynebacterium* spp., *Micrococcus* spp.

Of the Gram-stain variable microorganisms, *Corynebacterium* spp. belonging to a suborder of microorganisms within the *Actinomycetales* family, which are known for having a unique cell wall structure (13, 14). *Corynebacterium* spp. and others belonging to the suborder have an additional lipid bilayer (two in total) where the second lipid bilayer consists mainly of mycolic acid; long-chain α-alkyl, β-hydroxy fatty acids (15). While *Micrococcus* spp. does not have the second lipid bilayer and therefore has less phospholipids present in the cell wall structure. With the feature selection algorithm, 990-1005, 1057-1088, 1137-1169, 1180-1196, 1230-1383, 1385-1401, 1403-1416, 1437-1457, 1459-1470 cm^{-1} were selected for optimal discrimination between the two groups of spectra. It is noteworthy to mention through visual inspection of the averages of the

two groups of spectra, there were 3 distinct wavenumbers in the variance spectrum calculated from the spectral average of each of the two groups: 1085, 1380 and 1401 cm^{-1} (Figure 4.1).

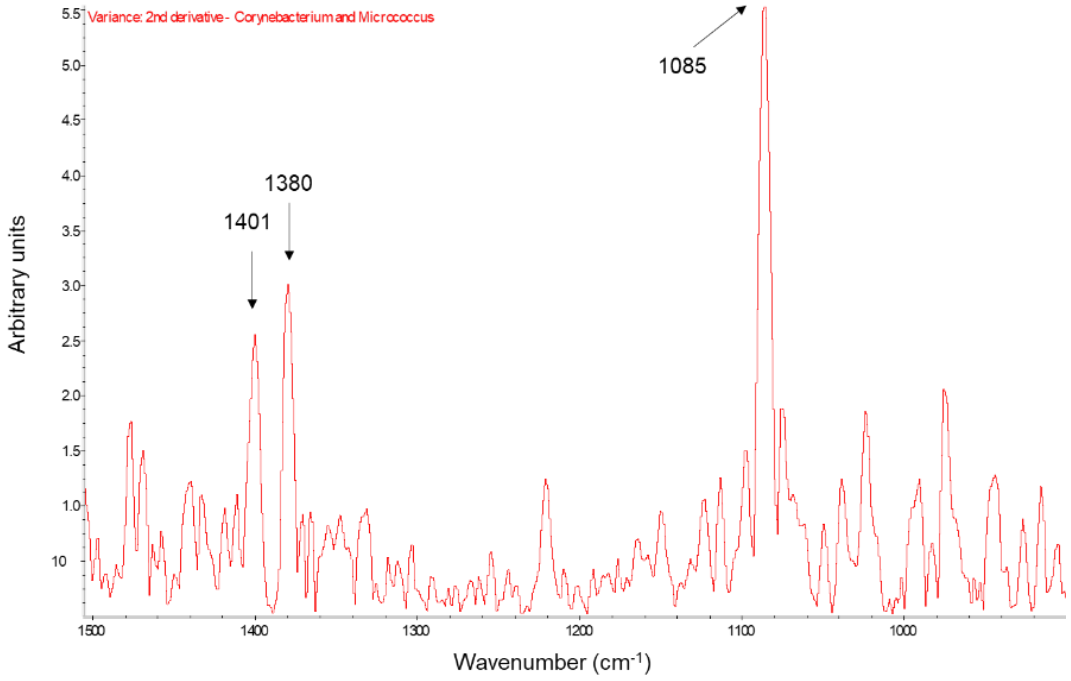


Figure 4.1. Variance spectrum of 2nd derivative spectra calculated from the spectral average of each of the two groups belonging to *Corynebacterium* spp. and *Micrococcus* spp.

These three peak positions are associated with the P=O symmetric stretching of phosphates (1085 cm^{-1}) and COO^- symmetric stretching associated with amino acid side chains and fatty acids (1380 and 1401 cm^{-1}) were among the selected wavenumbers found using the feature region selection algorithm. The intense spectral variance between *Corynebacterium* and *Micrococcus* spp. at 1085 cm^{-1} may be attributed from the second lipid bilayer present in *Corynebacterium* spp. and absent in *Micrococcus* spp.

To recap, the pairs located at a higher-level tier such as tier 4 are more spectrally similar compared to pairs at a lower tier level such as tier 1. The spectral variance between *Corynebacterium* spp. and *Micrococcus* spp. are much less than the spectral variance between tier 1 pairs such as the pair consisting of spectra acquired from bacteria and fungi (Figure 4.2).

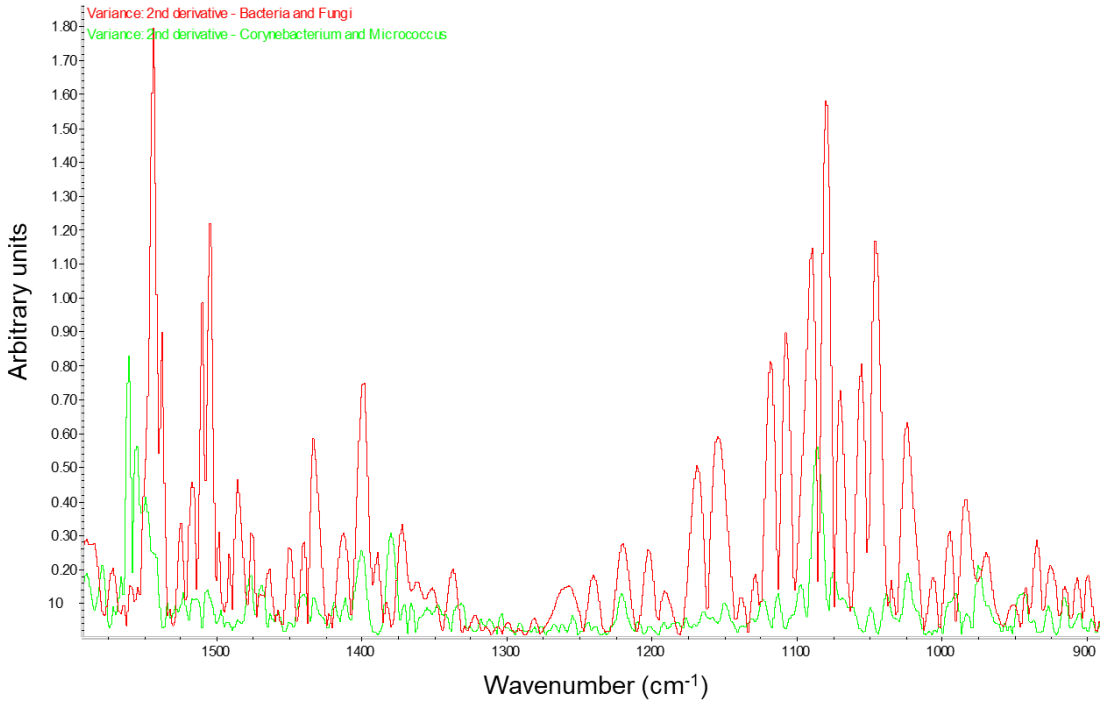


Figure 4.2. Variance spectra of 2nd derivative spectra of average spectra of bacteria and fungi spectra (red), and of average spectra of *Corynebacterium* spp. and *Micrococcus* spp. (green)

Set A of the spectral reference database was constructed to predict Gram-stain variable microorganisms (i.e., *Bacillus* spp., *Corynebacterium* spp. and *Micrococcus* spp.) from Gram-negative and Gram-positive bacteria. Additionally, the spectral reference database structure Set A also discriminates (screens out) spectra belonging to fungi before continuing to bacterial identification. More details on the construction of the yeast spectral database are described in CHAPTER 5. At the completion of Set A, Sets B and C spectral reference database structures are for the genus and species level for bacterial identification from spectra belonging to Gram-positive and Gram-negative bacteria respectively.

4.4.1.2. Set B: Construction of ATR-FTIR multitier pairwise spectral database for the discrimination of genus and species belonging to Gram-positive bacteria

Employing the same method of spectrotyping in Set A to Set B, spectra belonging to Gram-positive bacteria were separated into 8 pairs across 5 different tiers (Table 4.3 and Figure 4.15). The Set B model allows for the genus identification of *Enterococcus*

spp., *Staphylococcus* spp. *Streptococcus* spp. and *Listeria* spp. At a higher (more specific) taxonomic level such as the species level, Set B allows for species identification of *Enterococcus faecium*, *Enterococcus faecalis*, *Enterococcus gallinarum*, *S. aureus* and *Listeria monocytogenes*.

Table 4.3. Set B: Selected regions through a feature selection algorithm for discrimination between two pairs of spectral groups of Gram-positive bacteria for the construction of an ATR-FTIR spectral reference database for microorganism identification

Tier	Group	Pair	Region (cm^{-1})
1	a	*Enterococcus* spp. and *Streptococcus* spp.	1180-1185,1314-1319
	b	*Listeria* spp. and *Staphylococcus* spp.	
2	a	*E. faecalis*, *E. gallinarum* and *Streptococcus* spp.	980-1051,1053-1480
	b	*E. faecium*	
2	c	CoNS	999-1005,1025-1031,1064-1079,1362-1368,1469-1474
	d	*S. aureus* and *Listeria* spp.	
3	a	*Enterococcus* spp.	1090-1100,1146-1152,1271-1277,1340-1346,1349-1355
	b	*Streptococcus* spp.	
3	c	*Listeria* spp.	1096-1105,1236-1241,1364-1370
	d	*S. aureus*	
4	a	*E. faecalis*	1480-980
	b	*E. gallinarum*	
4	c	*Listeria monocytogenes*	990-995,1008-1029,1055-1070,1120-1133,1150-1155,1165-1174,1206-1221,1254-1267,1277-1293,1351-1357
	d	*Listeria monocytogenes* and non-*Listeria monocytogenes*	
5	a	*Listeria monocytogenes*	1310-1000
	b	non-*Listeria monocytogenes*	

Following the same workflow as Set A, the least spectral similarity is at a lower tier level and increasing in tier level will increase in spectral similarity. However, it should be noted that multiple genera can be grouped together, for example, *Enterococcus* spp. and *Streptococcus* spp. are grouped together before being discriminated; and they are being discriminated between *Staphylococcus* spp. and *Listeria* spp. Based on spectrotyping, it was found that the spectral similarity between the two pairs (pair 1: *Enterococcus* spp. and *Streptococcus* spp.; pair 2: *Staphylococcus* spp. and *Listeria* spp.) are more similar to than between the two pairs, as such, the two pairs make up the single pair at Tier 1 of Set B.

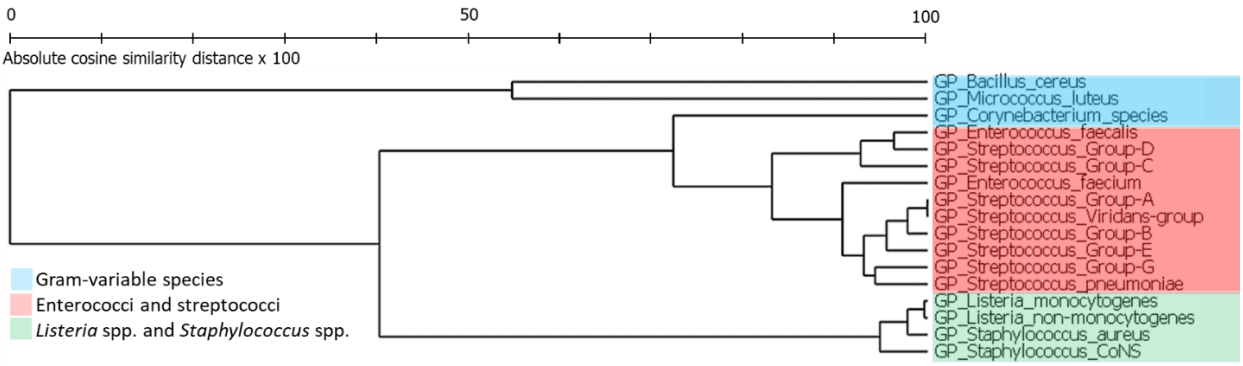


Figure 4.3. Dendrogram of unsupervised hierarchical cluster analysis of averaged reference ATR-FTIR spectra of Gram-positive bacteria using the spectral region between 1370 and 1000 cm^{-1}

Interestingly, *Enterococcus* was once classified as belonging to the streptococci group based on serological classification based on the Lancefield classification system, however, based on 16S rDNA sequencing and DNA-DNA/DNA-rDNA hybridization, the species were reclassified into the new *Enterococcus* spp. genus (16). With that note, it is interesting to observe similar clustering through HCA of ATR-FTIR spectra within the regions of 1370-1000 cm^{-1} (Figure 4.3). Figure 4.3 illustrates the HCA of the average spectra of those Gram-positive bacteria used to create the spectral reference database for microbial identification without a region selection, unsupervised spectrotyping of the Gram-positive bacteria resulted in *Enterococcus* and *Streptococcus* clustering in the same arm of the dendrogram - displaying high spectral similarity (>90%). Spectra collected from these two genera are therefore more related to the genotypic 16S rDNA and DNA-DNA/DNA-rDNA hybridization classification compared to their phenotypic serological classification.

As observed with *Enterococcus* and *Streptococcus* spp., spectrotyping for the selection of pairs for the construction of the spectral reference database is not limited to current nomenclature and/or taxa classification. At the end of Set B, spectra belonging to *L. monocytogenes* are broken up into two different pairs and tiers (Table 4.3 and Figure 4.15). Spectrally, *L. monocytogenes* isolates were clustering in two separate groups and therefore could not be grouped into one as observed for the other genera. *L. monocytogenes* spectra may be clustering into two separate spectral groups of

L. monocytogenes, due to differences in serovars and further classification of the strains would be needed, however, strain typing of the isolates were not completed for the current study. Other studies have however successfully classified serovars of *L. monocytogenes* by FTIR spectroscopy-based methods and analysis tools (17-19).

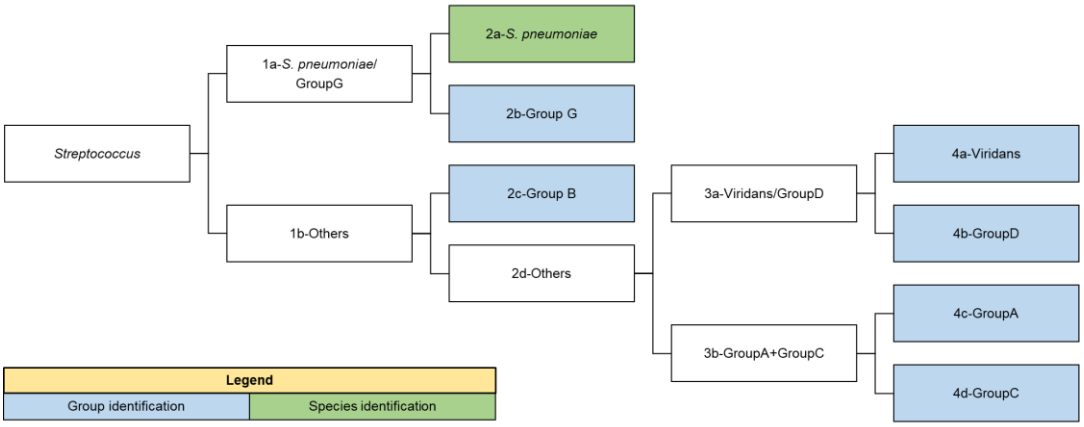


Figure 4.4. Spectrotyping as a method for constructing a pairwise multitier structure ATR-FTIR spectral search database for *Streptococcus* serotypes

Moreover, through spectrotyping of streptococci, capsular serotypes groups A, B, C, D and Viridans groups can be identified, however, species within the groups are not spectrally distinguishable through the outlined spectrotyping method (Figure 4.4). Similar capsular serotyping for *Streptococcus* results were also achieved by transmission mode FTIR spectral acquisition and HCA using the spectral region between 1300 and 800 cm^{-1} (20).

Current spectroscopy-based microorganism identification methods such as MALDI-TOF MS have reported 80-97% correct *Streptococcus* species identification, however, upon further investigation by mass peak analysis (not available for clinical routine use), higher correct identification rates were achieved (21). In the current study, there are limited representation of streptococci species per serogroup and may be the cause of unsuccessful species discrimination through spectrotyping, additional spectral data are required to fully assess the potential. The result of the current study suggests that spectroscopic analysis of streptococci species may be unreliable but for serotyping, may be more appropriate.

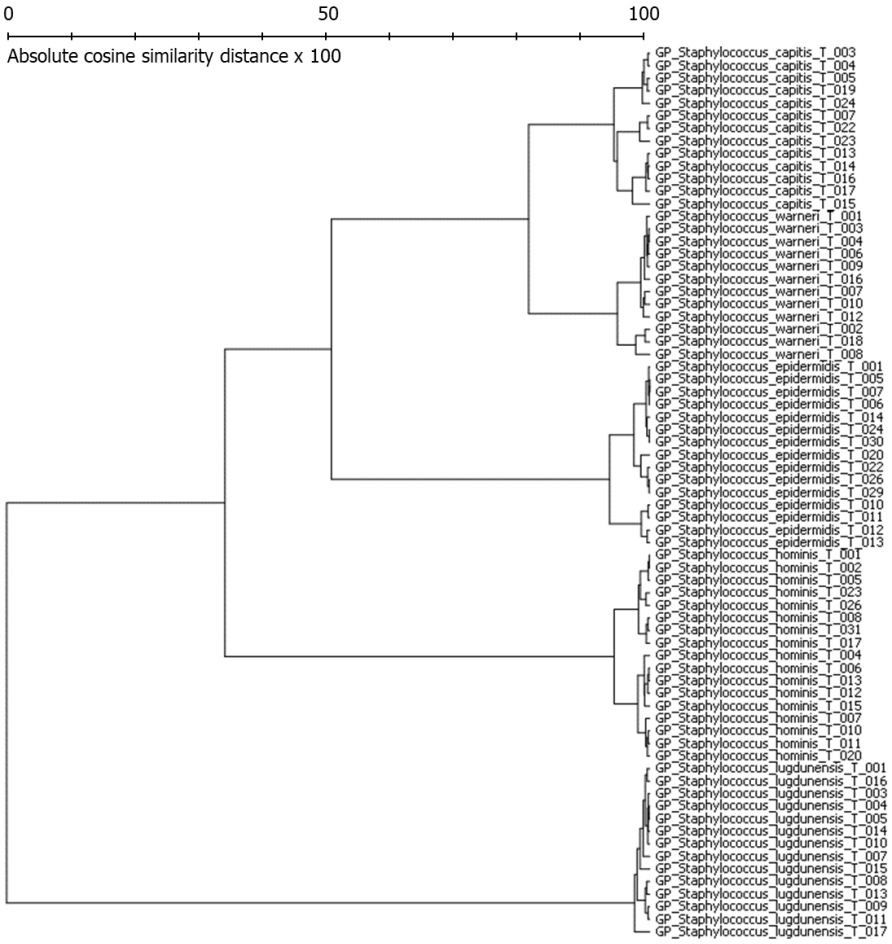


Figure 4.5. Feature selection and hierarchical cluster analysis using cosine similarity distance and ward linkage of ATR-FTIR spectra of 5 coagulase negative staphylococci species

Moving forward, although there were ATR-FTIR spectra collected from multiple species belonging to *Staphylococcus*, the spectral reference database structure does not include the staphylococci species other than *S. aureus*, which is the most clinically relevant *Staphylococcus* species. The remaining staphylococci species are grouped based on the absence of the coagulase enzyme and are classified/known as coagulase negative staphylococci (CoNS). Additionally, discrimination of staphylococci species by FTIR spectroscopy has already been explored and successfully demonstrated its ability to differentiate between 39 difference species (e.g., *Staphylococcus capitis, Staphylococcus caprae, Staphylococcus cohnii, Staphylococcus epidermis, Staphylococcus hominis, Staphylococcus haemolyticus, Staphylococcus warneri*, etc.) (22). Furthermore, species within CoNS can be discriminated from one another according to Guliev's group without

the need for the outlined pairwise multitier approach. Guliev's group collected FTIR spectra in transmission mode and processed the microbial samples by inactivating the microorganisms in ethanol. The current work re-valuated the previously mentioned study to apply it to the proposed whole organism ATR-FTIR microbial identification technique (without the need for inactivation). Utilizing the feature selection algorithm on a selected few CoNS species, *S. capitis*, *S. epidermidis*, *S. hominis*, *Staphylococcus lugdunensis* and *S. warneri* (opposed to selecting pairs of CoNS species for pairwise multitier discrimination), a dendrogram was generated by HCA (Figure 4.5). The feature selection region selection and HCA resulted in successful discrimination between the selected CoNS species, confirming that the ATR-FTIR spectra without ethanol treatment and drying, was also feasible for complete discrimination between of selected species. Moreover, Figure 4.5 illustrates that *S. capitis* and *S. warneri* are spectrally the most similar relative to the other species, then comes *S. epidermidis*, *S. hominis* then *S. lugdunensis* in decreasing similarity.

Multiple studies have evaluated conventional biochemical (or phenotypic) methods for CoNS species identification using commercial systems such as API® ID 32 STAPH (bioMérieux, Marcy-l'Étoile), VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile), MicroScan Pos ID (Baxter Diagnostics Inc., CA) and Phoenix 100 ID/AST (Becton Dickinson, MD) which produced a wide variety of accuracy performances of as low as 75.2% to as high as 96.8% correct species identification (23-27). With CoNS increasing in prevalence in hospital acquired infections, adequate identification methods are necessary for appropriate diagnostic and treatment (28) and information in the ATR-FTIR spectra may contribute to advancing rapid microorganism identification.

Today, phenotyping methods are complimented with genotyping methods to reassign species designation based on genetic relatedness (29). Genotypic methods have however, been used to identify CoNS species using methods such as PCR using sequences obtained from selected DNA target sites such as 16S ribosomal DNA (rDNA). Although 16S rDNA have been successful for species identification of various microorganisms such as *Brucella* spp. and *Acinetobacter* spp. (30, 31) and has been used to study bacterial genetic relatedness (32), for species belonging to the CoNS group,

some of the species may have identical 16S rDNA sequences and in turn would not be applicable for species identification. As such, PCR sequencing of other genes such as *tuf* and *sodA* have been explored (23, 33-35). The *sodA* gene have been previously reported to be responsible for encoding for the manganese-dependent superoxide dismutase in Gram-positive cocci (33, 36, 37). While PCR gene sequencing of the *sodA* gene of various strains of CoNS has provided researchers with the relationship between the different strains through a phylogeny tree.

In two studies performing PCR gene sequence of the *sodA* gene of *S. capitis*, *S. epidermidis*, *S. hominis* and *S. lugdunensis*, based on the obtained phylogeny tree, reported agreeance with *S. capitis* being closely related to *S. warneri* followed by *S. epidermidis*, *S. hominis* and lastly, *S. lugdunensis* (33, 35). Moreover, ATR-FTIR spectra of those mentioned CoNS species, using the feature region selection algorithm and HCA, also resulted in similar trends using spectral data (Figure 4.5). Based on the cosine similarity distance between the spectra of the 5 CoNS species, the ATR-FTIR spectra has some correlation with the expression of the *sodA* gene. This example further describes the simplicity of datamining the complex information found in ATR-FTIR spectra. Not only can the spectrotyping method be utilized for creating ATR-FTIR spectral references databases (multitier pairwise), but it can also be used to potentially strain type microorganisms.

4.4.1.3. Set C1: Gram-negative ATR-FTIR multitier pairwise spectral reference database construction

Set C deals with pairs of spectra belonging to Gram-negative (GN) bacteria separated into Set C1 (spectrotyping set) and Set C2 (SVM set). Unlike Gram-positive (GP) bacteria, spectra belonging to GN bacteria proved to have more difficulties with spectrotyping for discriminating between the species for identification and required additional powerful analysis tools such as SVM. Again, Set C follows the same logic as Sets A and B where increasing tier levels are associated with increasing spectral similarity between spectral groups.

Table 4.4. Set C1: Selected regions through a feature selection algorithm for discrimination between two pairs of spectral groups of Gram-negative bacteria for the construction of an ATR-FTIR spectral reference database for microorganism identification

Tier	Group	Pair	Region (cm^{-1})
1	a	Remaining GN bacteria[1]	1006-1012,1049-1055,1148-1154,1239-1245,1277-1282,1295-1301,1303-1314,1400-1405
	b	Mucoid *Pseudomonas aeruginosa*	
2	a	*Burkholderia gladioli* and *Burkholderia cepacia complex* (BCC)	1318-1174
	b	Remaining GN	
3	a	*Burkholderia gladioli*	1001-1006,1021-1034,1226-1237,1277-1282,1349-1355
	b[2]	BCC	
3	c	*Acinetobacter* spp. + non-mucoid *P. aeruginosa* and *Enterobacteriaceae*[3]	1081-1087,1146-1152,1165-1176,1273-1278,1290-1303
	d[2]	NFGNB (*Stenotrophomonas* spp., *Achromobacter* spp.) and *Kingella* spp., *S. sonnei*	
4	c	*Acinetobacter*	999-1005,1098-1103,1306-1312,1381-1387,1463-1469
	d[2]	*Enterobacteriaceae*[3] and *P. aeruginosa*	

[1]Reamaining GN bacteria refers to all genera and species in the table that are not mucoid *P. aeruginosa*

[2]Group of spectra used for SVM (difficulties to spectrotype)

[3]*Enterobacteriaceae*: *Escherichia*, *Shigella*, *Salmonella*, *Citrobacter*, *Enterobacter*, *Klebsiella*

A total of 5 pairs were selected spanning over 4 tiers, however, through spectrotyping, only mucoid *P. aeruginosa*, *Burkholderia gladioli*, *Burkholderia cepacia* complex (BCC) and *Acinetobacter* spp. were identified, and the remaining genera and species were reserved for Set C2 for analysis with SVM (Table 4.4).

Upon group selection for a pair at the start of Set C1, a group of spectra belonging to *P. aeruginosa* were significantly dissimilar from all the other spectra. The dissimilar *P. aeruginosa* is considered group 1 while the other GN bacteria (also including some spectra belonging to *P. aeruginosa*) is considered group 2. Group 1 spectra of *P. aeruginosa* without preprocessing demonstrates lower absorbances between 3000-2836 cm^{-1} and 1576-950 cm^{-1}. These regions are associated with all biomolecules within the microorganism and with a common scale, the average spectrum of Group 1 (lower absorbance signals) displays lower signals in those mentioned regions (Figure 4.6). This

is therefore associated with lower biomass and was confirmed with additional strain information. Those spectra belonging in Group 1 were collected from *P. aeruginosa* grown on blood agar with a mucoid phenotype coming from children diagnosed with CF.

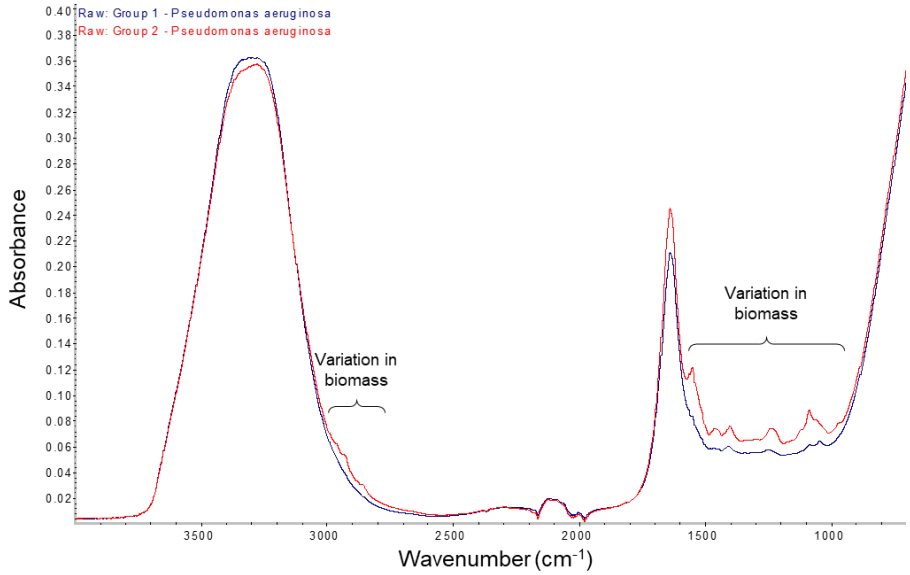


Figure 4.6. Superimposed spectra of unprocessed average spectra of Group 1 *Pseudomonas aeruginosa* (mucoid) and Group 2 *P. aeruginosa* (non-mucoid)

Mucoid strains of *P. aeruginosa* are due to the increased synthesis of alginate exopolysaccharides due to the overexpression of the *algD* gene among others. The *algD* gene encodes for an enzyme (guanosine diphosphate mannose dehydrogenase) which acts as a catalyst in the synthesis process of alginate precursors (38). Interestingly, mucoid *P. aeruginosa* have the ability to revert to non-mucoid phenotype and back to mucoid spontaneously, indicating environmental factors are responsible for the expression of *algD* gene and that these strains are the same and are not mutants, but are variants (39). Studies have determined that the extracellular polymeric substance (EPS) of *P. aeruginosa* are mainly composed of different polysaccharides [alginate, Psl (mainly of mannose and galactose polysaccharides) and Pel (glucose-rich polysaccharides), proteins, lipids, and DNA (with varying proportions) (40-44). Observing the only spectral variation between average spectra of non-mucoid *P. aeruginosa* and mucoid *P. aeruginosa*, there are high spectral variances in the lipid associate region (C-H stretching of lipids at 3000-2800 cm^{-1}) predominantly at 2923 and 2852 cm^{-1}; protein region with the amide II band at 1543 cm^{-1} being the highest variance band followed by high variances

at 1085 cm^{-1} which is associated with carbohydrates and/or phospholipids (Figure 4.7). As such, this suggests the difference in protein and lipid profiles between the two groups of average spectra. Moreover, the lack of biomass absorbance may be due to increase is mucoidal material which is of higher content than the biomass of the bacteria and results in relative weaker infrared absorbances.

With the major spectral differences between mucoid *P. aeruginosa* and the remaining GN bacteria (including non-mucoid *P. aeruginosa*), discrimination between the two groups of spectra is first in the GN Set C1 of the spectral reference database (Table 4.4).

Unlike Sets A and B, Set C groups of spectra are less related to current taxonomy of the microorganisms. For example, NFGNB (*Acinetobacter*, *Bordetella*, *Burkholderia*, *Legionella*, *Moraxella*, *Pseudomonas* and *Stenotrophomonas*) are a group of bacteria that are unable to catabolize glucose and therefore have the inability to ferment sugars and do not belong to the *Enterobacteriaceae* family. NFGNB, similar to *Enterobacteriaceae*, are also challenging to identify and are intrinsically resistant to antibiotics (45). Although, *Pseudomonas* spp. are classified as NFGNB, they show species diversity and multiple strain variants which are observed in the ATR-FTIR spectra and are grouped into multiple pairs (Table 4.4).

Moreover, NFGNB are slow growers and are biochemically weak (low biochemical activity during biochemical assimilation tests) making the use of conventional media unsuitable for identification (46). These microorganisms are found mainly in the environment and cause infections to those who are immunocompromised, such as those diagnosed with CF. NFGNB also lack distinct phenotypic characteristic for differentiating between one NFGNB species to another (e.g., isolates being identified as *Stenotrophomonas* by conventional phenotypic methods while by partial 16S rRNA sequencing suggests genus identification of *Pseudomonas* (47)). Current methods of identifying NFGNB include colony morphology, glucose fermentation, carbon source, and gelatin hydrolysis. These identification methods may be inconclusive, unreliable, time-consuming and require up to several days for results (4).

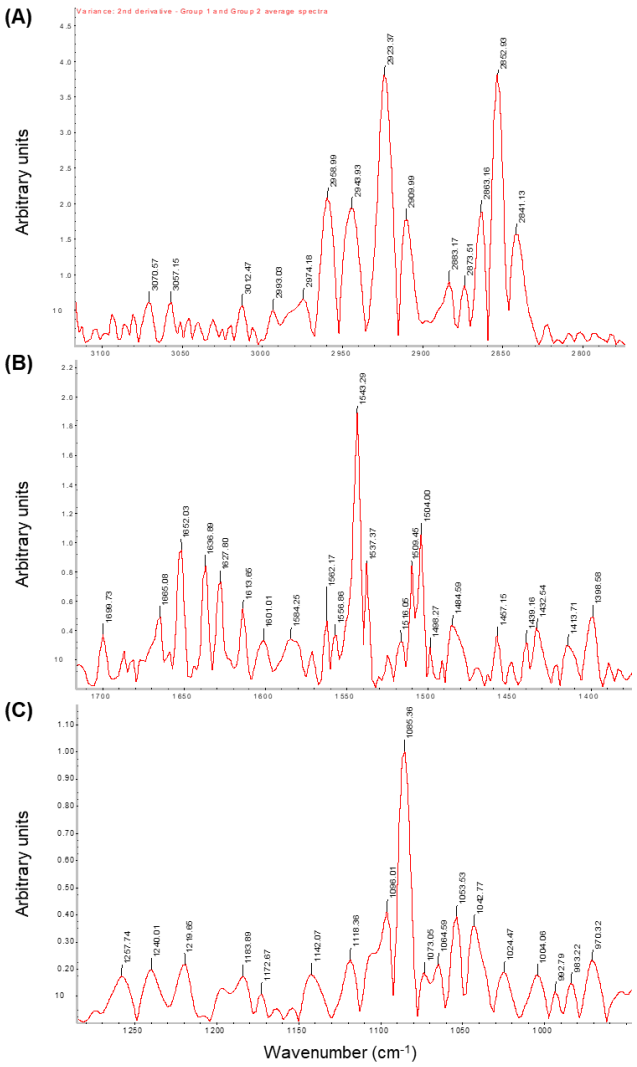


Figure 4.7. Variance spectrum of 2nd derivative, vector normalized and averaged Group 1 (mucoid *Pseudomonas aeruginosa*) and Group 2 (non-mucoid *P. aeruginosa*) in regions (A) 3100-2800 cm^{-1} (B) 1700-1350 cm^{-1} and (C) 1275-900 cm^{-1}

In addition, the biochemical activity of the NFGNB are low where commercial identification methods are reported to produce misidentifications (5). Although they are biochemically weak, the differences between the NFGNB microorganisms are spectrally dissimilar from each other and spectrotyping is achievable for those genera (Figure 4.8).

While discrimination between CF-related NFGNB is possible with HCA and region selection relative to each other (Figure 4.8), spectra belonging to NFGNB bacteria combined with the other GN bacteria spectra, pairwise selection of groups were not achievable through spectrotyping – resulting in similar difficulties for identification as

current biochemical methods. NFGNB groups of spectra had to be separated into various tiers and pairs for the construction of the spectral database. With region selection, *Burkholderia* spp. (species within BCC and *B. gladioli*) separation between the two groups were possible within 1318-1174 cm^{-1}, which are associated with amide III band components of proteins, P=O stretching (symmetrical and asymmetrical) of phospholipids and phosphodiesters respectively and various vibrations associated with carbohydrates in the cell wall and differences in glycosidic linkage configurations (48-51).

It is interesting to note that *Burkholderia* spp. are well-known for developing biofilms with EPS. The polysaccharides within the biofilms (cell-bounded or un-bounded) have been widely studied and characterized; containing galactose, glucose, mannose, rhamnose and glucuronic acid; however, EPS also does contain DNA, proteins and lipids (52-54).

Unlike the other GN bacteria present in the spectral reference database, *Burkholderia* spp.'s biofilm and biochemical composition are significantly dissimilar from the others and are observed in the ATR-FTIR spectra. Additionally, the NFGNB species (excluding *P. aeruginosa*) are spectrally similar relative to spectra belonging to those species in the *Enterobacteriaceae* family.

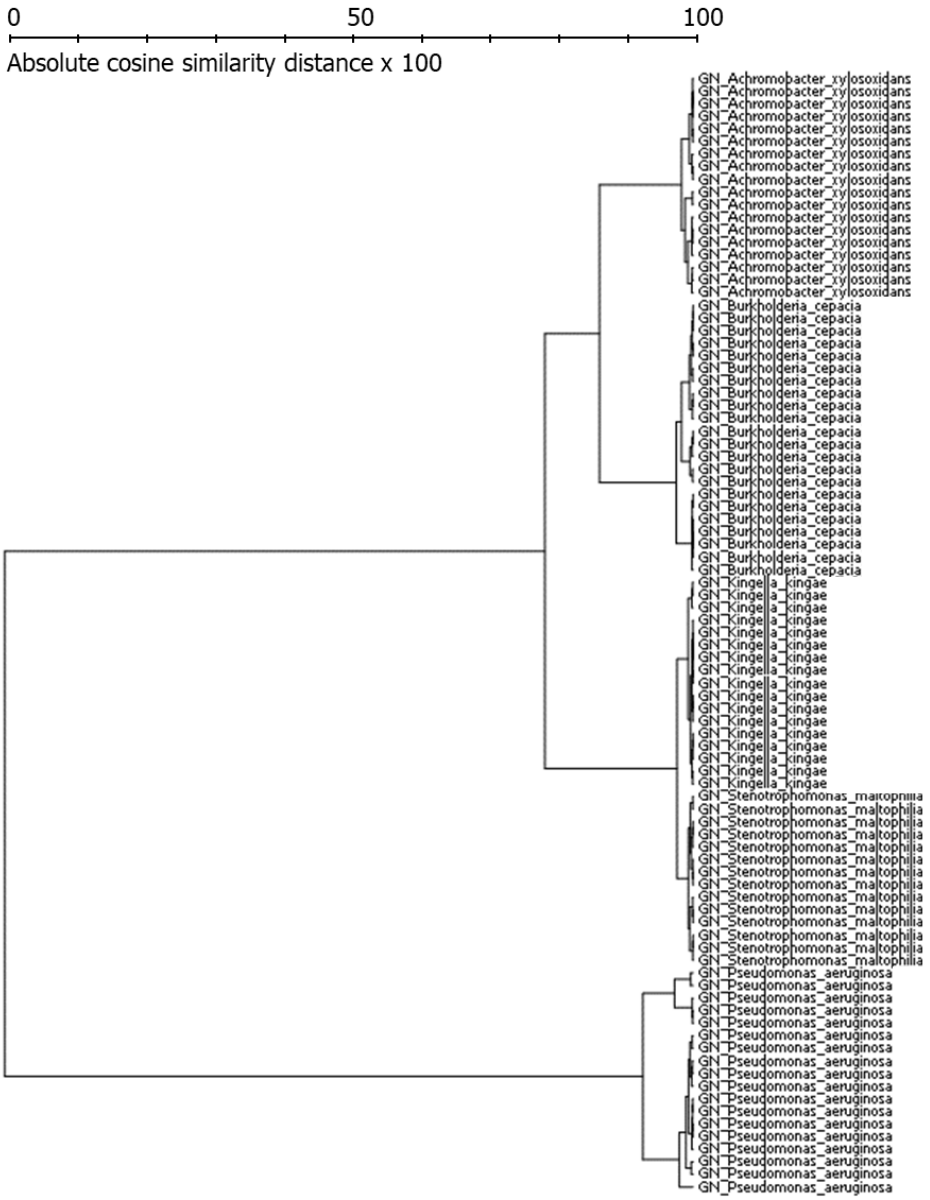


Figure 4.8. Dendrogram illustrating complete discrimination of cystic fibrosis related non-fermenting Gram-negative bacteria based on hierarchical cluster analysis and region selection (region: 1051-1057, 1146-1157, 1159-1165, 1167-1174, 1182-1187, 1275-1293, 1336-1342, 1346-1353, 1355-1360, 1450-1456 cm^{-1})

Other difficult to identify microorganisms in routine also include species within the *Enterobacteriaceae* family (e.g., *Escherichia*, *Shigella*, *Salmonella*, *Citrobacter*, *Enterobacter*, *Klebsiella*). The difficulty to identify microorganisms arises from their high similarities in their 16S rRNA genes (>99.5% relatedness) but low relatedness (23-50%)

from their whole genome sequence (32). Carbon assimilated tests demonstrates difficulties between various genera and species of the *Enterobacteriaceae* family (59).

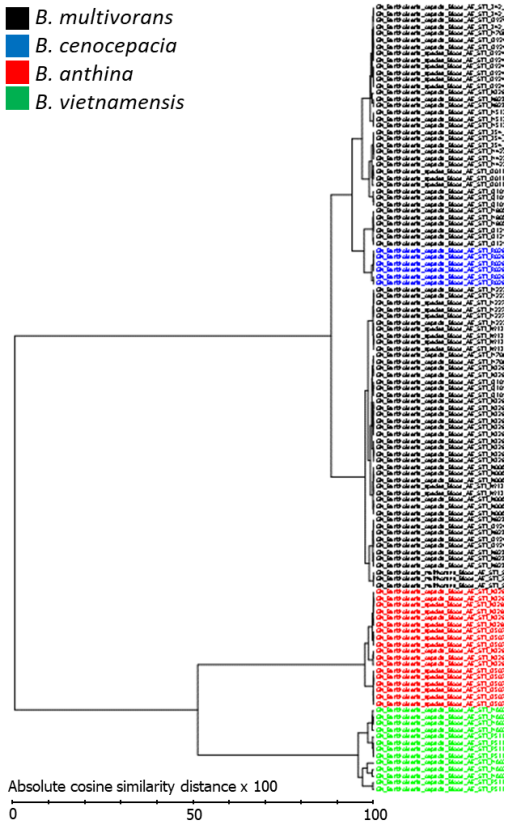


Figure 4.9. Dendrogram generated from hierarchical cluster analysis for the discrimination of *Burkholderia cepacia* complex species relating to cystic fibrosis infections; *B. multivorans*, *B. cenocepacia*, *B. anthina* and *B. vietnamensis*

Aside from conventional biochemical identification techniques, 16S rRNA gene sequencing have been widely accepted for species classification and identification but have difficulties differentiating between phylogenetically similar genera including those belonging to the *Enterobacteriaceae* family with *Enterobacter* being one of the most difficult to differentiate species (32, 60, 61). Currently, MALDI-TOF MS has been widely used to rapidly identify GN bacteria, overcoming the difficulties found in gene sequencing (62). As such, spectrotyping may be the linkage between both phenotypic and genotypic classification and be used as a tool to aid classification of microorganisms for the sole intent of identification.

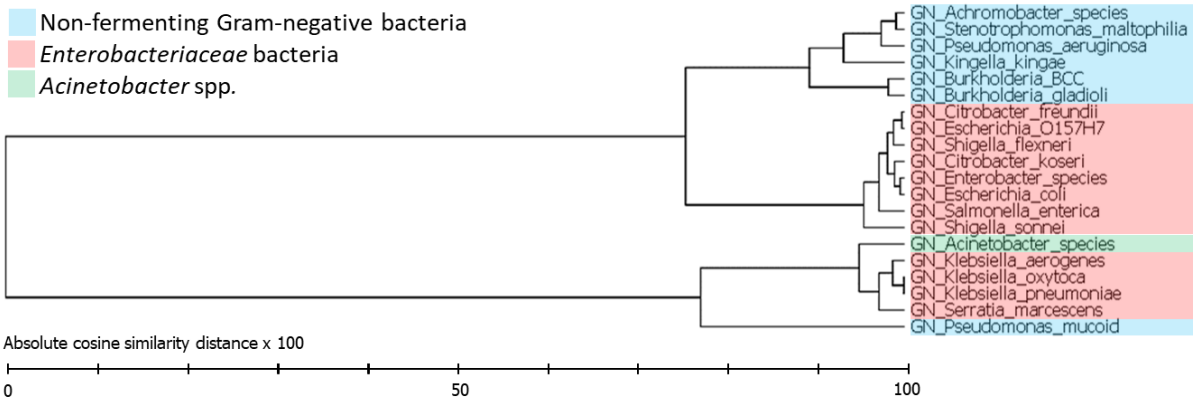


Figure 4.10. Hierarchical cluster analysis of ATR-FTIR spectra of averages of Gram-negative bacteria depicted using relative cosine similarity distance, ward linkage and a broad spectral region of 1660-900 cm^{-1}

For the construction of the ATR-FTIR spectral reference database, discrimination between spectra belonging to strains of the *Enterobacteriaceae* family proved to be challenging due to the high spectral similarity between *E. coli*, *Citrobacter* spp. and *Enterobacter*. The proposed spectrotyping method has limitations when it comes to GN bacteria, which may be due to the thinner cell wall compared to GP bacteria where there are less IR absorption signals and low spectral variances between species. As such, the feature selection algorithm is not powerful enough for creating a model for GN species identification. Although the ability to select pairs is limited with the outlined model to construct the multitier pairwise spectral database, the method and algorithm have proven successful in various levels of strain-typing within a closed model such as, species discrimination of CoNS species (Figure 4.5), BCC species (Figure 4.9), *E. coli* O157 versus non-*E. coli* O157:H7 (Figure A.7), serogroup discrimination of *Salmonella enterica* (Figure 4.13) and discriminating between *L. monocytogenes* versus non-*L. monocytogenes* (Table 4.3) and *Shigella* versus *E. coli* (Figure A.6). The feature selection algorithm is not limited to the suggested multitier pairwise database structuring method, but it can also be used to create single models for pairwise discrimination.

Looking closely at the averaged spectra of all GN bacteria presented in the database and performing unsupervised HCA (Figure 4.3), although phylogenetically similar, the spectral distances between *E. coli* and *Shigella flexneri* were smaller than the spectral distance to

Shigella sonnei. The larger spectral distance from *S. sonnei* from the other two may be attributed from their differences in cell wall structure and configuration of the lipopolysaccharide (LPS) O antigens, where *Shigella* and *E. coli* share multiple O-antigens except for *S. sonnei* which only has 1 O-antigen (63-65). As such, the major differences in cell wall structure are displayed in their ATR-FTIR spectra, which does not relate to their ancestry. ATR-FTIR spectrotyping may therefore be argued to be more related to phenotypic attributes rather than genotypic variability.

While species discrimination (e.g., CoNS and BCC species) through non-pairwise spectrotyping, pairwise is optimal for the reduction of interference of overlapping of spectral data when creating a search database for microorganism identification (Figure A.4). The discrimination between non-O157:H7 *E. coli*, *E. coli* O157:H7, *S. sonnei* and *S. flexneri* using the feature selection algorithm is achievable (Figure A.6), however, increasing the spectral data to increase spectral representation of the groups of spectra using the same spectral region will then increase the data points and display overlaps between the 4 groups of spectra (Figure A.4). Such that when all 4 groups are analyzed with HCA and the feature selection algorithm, there is at least 45% spectral similarity between 2 clusters (Figure A.5). However, reducing the 4 groups to only two groups of spectra based on genera (*E. coli* and *Shigella*), the spectral similarity increased to 67% (Figure A.6), becoming more specific. As such, multiclass spectrotyping may be utilized to discriminate between difficult to identify microorganism based on strain, serotype and others while pairwise discrimination is optimal for rapid screening of microorganisms.

Towards species discrimination, known differences between *S. sonnei* and *S. flexneri* are associated with the differences between their cell wall polysaccharide composition. The major difference between two species is the O-polysaccharide side chain of the LPS where *S. sonnei* has a small chain and *S. flexneri* has a long repeating chain (66). Unlike *S. sonnei*, *S. flexneri* has glycogen synthase (encoded by the gene *glgA*) that is responsible for the production of α-1,4-glucan chains from glucose (of the cell wall) (67).

The O-antigen is the major component of the LPS of the cell wall and is responsible for cell surface interaction (responsible for resistance to antibiotics) and accounts for 13% of

the dry cell weight. The O-antigen within species of *Shigella* displays enough structural diversity and functionality for serotyping (68).

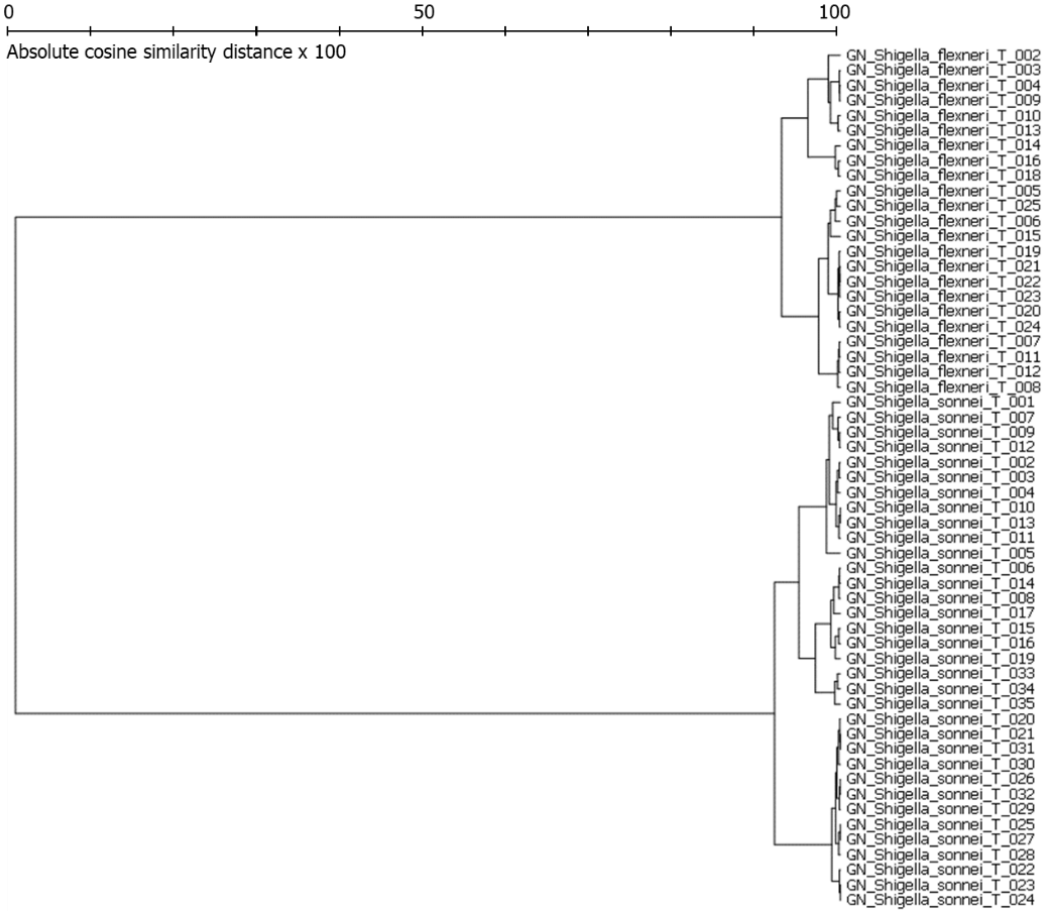


Figure 4.11. Dendrogram generated from unsupervised hierarchical cluster analysis of ATR-FTIR spectra of *Shigella sonnei* and *Shigella flexneri* using a broad region of 1480-980 cm^{-1}

As such, the O-antigen in *S. flexneri* and *S. sonnei* have been reported to be extremely dissimilar from one another and may be the cause for effective discrimination between the species by ATR-FTIR spectroscopy resulting from significant spectral difference and significant variance between the spectra from the two species (Figure 4.11). It is also of interest to note that *Shigella* are non-lactose fermenting, however, only *S. sonnei* has the 3 genes responsible for lactose fermentation, but it lacks permease activity and therefore does not ferment lactose (69). Major phenotypic and genotypic differences between *S. sonnei* and *S. flexneri* have been well documented, however, the biomolecules

associated with differences in the spectral profiles of each species are undocumented and should be further investigated.

For example, a major spectral difference at 1022-1024 cm^{-1} may be directly related to the longer LPS O-polysaccharide side chain (associated with the O-C and C-O vibrations specific to cell wall structure) relative to *S. sonnei* (Figure 4.12).

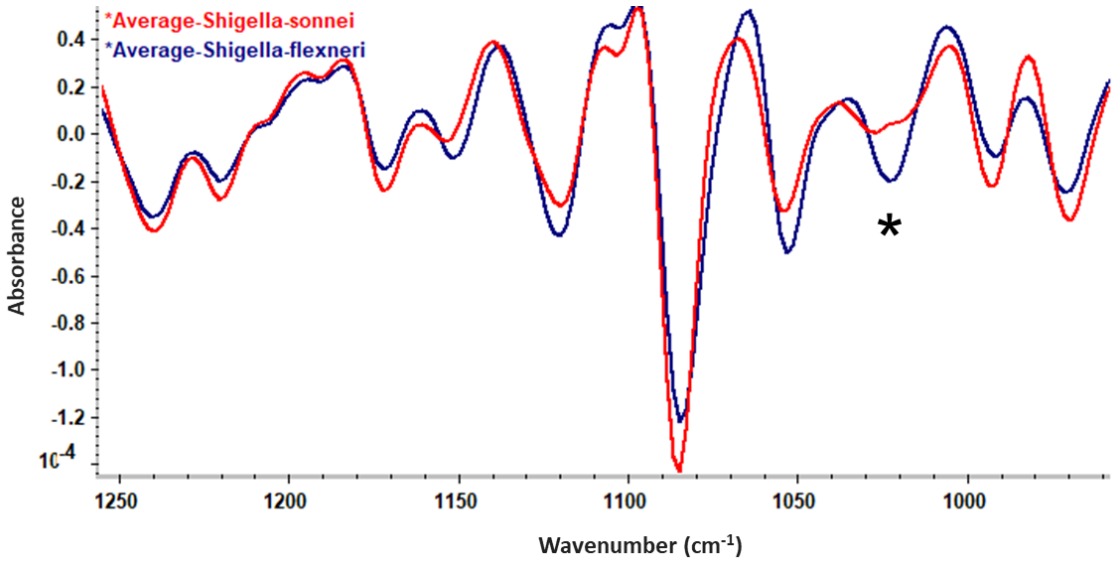


Figure 4.12. Superimposed 2nd derivative and vector normalized averaged ATR-FTIR spectra of *Shigella sonnei* and *Shigella flexneri* displayed in the spectral region of 1250-900 cm^{-1}. *Marks the 1022-1024 cm^{-1} regions possibly associated with carbohydrates of the lipopolysaccharide

Shigella and *Salmonella* are almost always pathogenic (with exception to *Salmonella bongori*) unlike the other genera of the same family (*Enterobacteriaceae*), as such it is extremely important to rapidly screen for these two microorganisms (70). Over the course of the study, only *S. enterica* were spectrally acquired and available and it is the most common and important species (of two species – *S. enterica* and *Salmonella bongori*). With limited species collected, for the construction of the ATR-FTIR spectral database, *Salmonella* is only identified to the genus level. Clinically, *Salmonella* species are classified based on their serovar and was investigated in the current chapter.

Figure 4.13 illustrates the results obtained from ATR-FTIR spectral analysis by HCA for *Salmonella* serogroup typing based solely on the differences in their ATR-FTIR spectra.

S. enterica Enteritidis and Thomson serovars belonging to serogroups D1 and C1 were effectively separated, while the serovars within serogroups B and C2 clustered together.

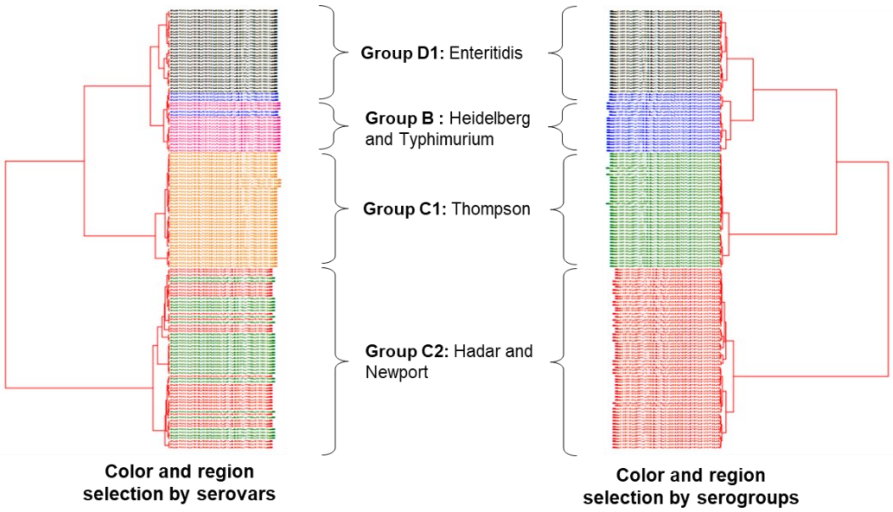


Figure 4.13. Dendrogram of *Salmonella enterica* serovar and serogroup discrimination through HCA and region selection of the top 6 serovars collected from food and animal feed in Canada

Serotyping is based on the O-antigen of the LPS and constitute a large portion of the bacterial cell wall. In this feasibility study, ATR-FTIR spectroscopy can distinguish groups of serovars with common antigens but are not able to distinguish biochemical differences between serovars of the same serogroup. The differences between serovars are based on the H-antigen (protein content of the flagella) composition and through the outlined spectrotyping technique, was unable to detect the differences in H-antigen variation within the serogroup.

Serogroups are referred to as the groups of O-antigens that are present and are further divided into serovars which are determined by the H-antigens. The ATR-FTIR spectra has the ability, under supervised conditions, discriminate between serogroups, however, does not have the ability to differentiate between the serovars. This suggests that for *Salmonella* serovars, spectral information using the current parameters are insufficient, however, based on the O antigen repeating polysaccharides units (commonly glucose, galactose, rhamnose and mannose (72)) is adequate for discriminating by the serogroups of *Salmonella*. Using the feature selection algorithm solely on the serovars of *Salmonella*

serogroup C2, regions were obtained and through HCA and PCA (Figure 4.11), discrimination between the two serovars were negative, further confirming lack of spectral information for the discrimination of *Salmonella* serovars.

Table 4.5. *Salmonella* serogroup, serovar and relating O antigen polysaccharide repeating unit structure

Serogroup	Serovar	O antigen polysaccharide repeating unit
B1	Heidelberg and Typhimurium	α1, 2[D-Man*-(α1→4)-L-Rha-(α1→3)-D-Gal] *Abe(α1→3)
C1	Thompson	α1, 2[D-Man-(β1→2)-D-Man-(α1→2)-D-Man-(β1→3)]
C2	Hadar and Newport	α1,4[L-Rha*-(α1→2)-L-Man-(α1→2)-D-Man-(α1→3)-D-Gal] * Abe(α1→3)
D1	Enteritidis	α1, 2[D-Man*-(α1→4)-L-Rha-(α1→3)-D-Gal] *Tyv(α1→3)

Man: mannose; Rha: rhamnose; Gal: galactose; Abe: abequose; Tyv: Tyvelose. Table adapted from (71) (72).

After closer examination, the spectral differences between the different serogroups may likely be attributed to the difference in LPS O-antigen polysaccharide repeating units where serogroups B and D1 has a difference of 1 carbohydrate and linkage (Table 4.3). What is more, serogroup C1 is more different than the other 3 serogroups with differences in linkages and carbohydrates. The cosine similarity distances in Figure 4.13 also ranked spectral similarity of the serogroup's similarity to the similarities of the polysaccharide repeating units of the O-antigen.

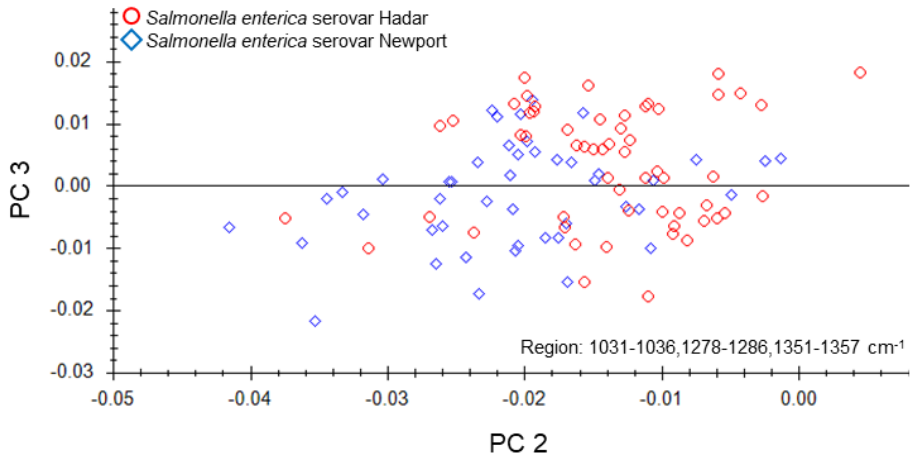


Figure 4.14. Feature selection and principal component analysis score plot for the discrimination of *Salmonella* serogroup C2 serovars Hadar and Newport

Although the outlined spectrotyping method is insufficient as a complete microorganism identification tool, combining other powerful tools such as SVM is another strategy that was examined. Through the multitier approach, Gram-stain variable, GN, GP and fungi were identified in Set A, then Set B identified various GP bacteria such as *S. aureus*, *E. faecalis*, *E. faecium*, *Streptococcus* species and the CoNS group. Lastly, Set C is incomplete and only mucoid *P. aeruginosa*, *B. gladioli* and *Acinetobacter* spp. where identified. *Burkholderia* species belonging to the BCC were also identified, but only to the complex level. While Set C1 is incomplete, through the spectrotyping method, individual models were studied and discrimination of species (e.g., BCC species, discrimination of *E. coli* and *Shigella*, CoNS species), serotype (e.g., discrimination between *E. coli* O157:H7 and non-O157 *E. coli*) and serogroups (e.g., *Salmonella* serogroups) were achievable. As such, there are significant information for discrimination in the ATR-FTIR spectra, however, in combination between other spectra belonging to other microorganisms, the model becomes weak with overlapping spectral data. The remaining species: *Burkholderia ambifaria*, *Burkholderia anthina*, *Burkholderia cenocepacia*, *B. cepacia*, *Burkholderia multivorans*, *Burkholderia vietnamensis*, non-mucoid *P. aeruginosa*, *S. maltophilia*, *Achromobacter* spp., *Kingella kingae*, *S. sonnei*, *S. flexneri*, *E. coli*, *Citrobacter* spp., *Enterobacter* spp., *S. enterica*, *Klebsiella* spp., and *Serratia* spp. require additional methods for identification using ATR-FTIR spectral data. To complete the identification model, SVM is introduced in Set C2 to complete the identification of the above listed microorganisms and the ATR-FTIR spectrotyping-based strategy pairwise multitier spectral database is summarized below in Figure 4.15 for reference of the work.

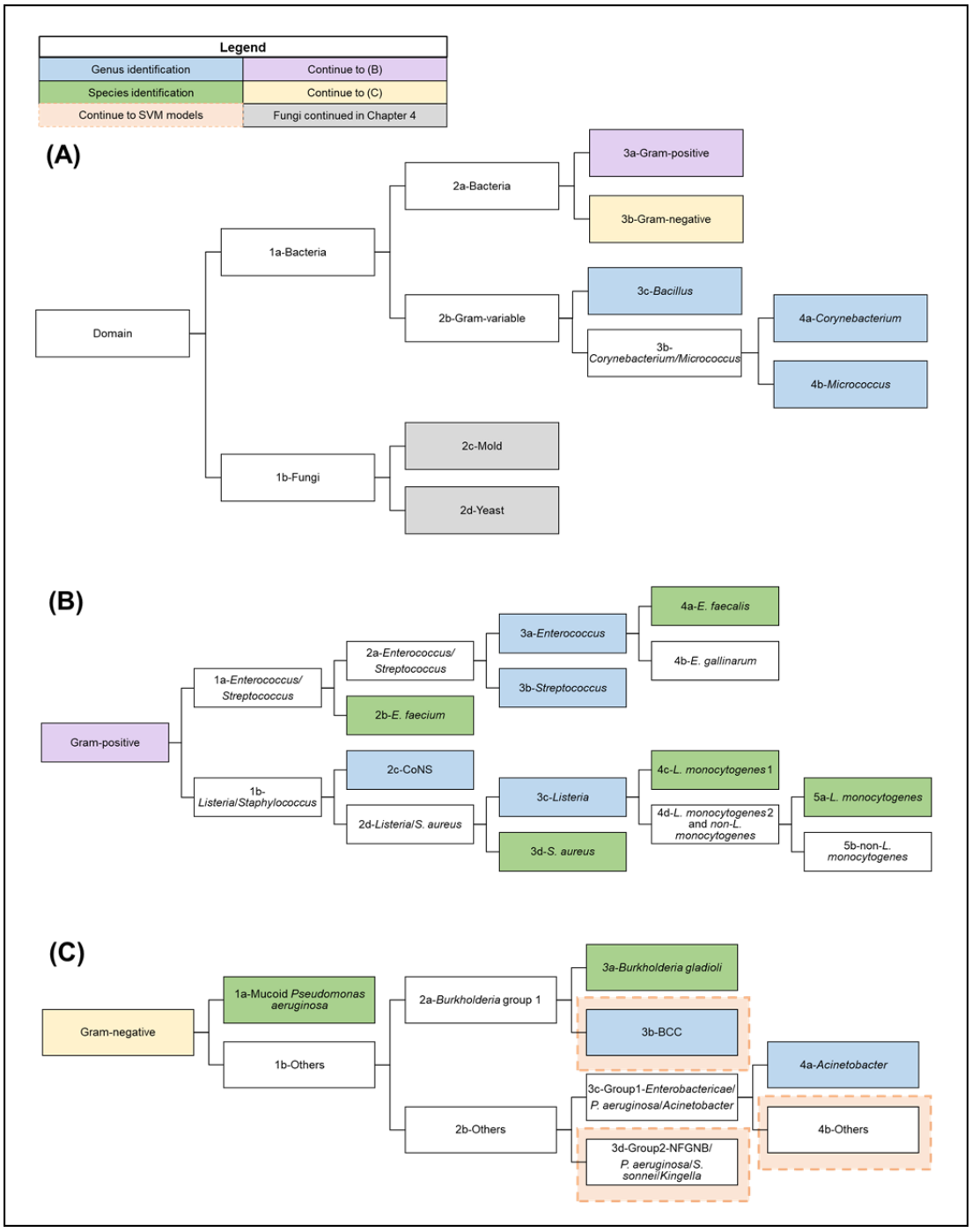


Figure 4.15. Summary of ATR-FTIR spectral reference database constructed structure for bacterial identification. (A) Set A pairwise structure for classifying Gram-stain type such as Gram-positive, Gram-negative and Gram-stain variable bacteria. (B) depicts the spectral reference database structure for Gram-positive bacteria and (C) or Set C1 illustrates the spectral reference database for Gram-negative classification and highlighting groups of spectra for SVM algorithms to classify the remaining Gram-negative bacteria.

4.4.1.4. Set C2: SVM models for continual GN bacteria classification by the ATR-FTIR-based technique for microorganism identification

Set C2 is the continuation of Set C1 where the spectrotyping technique was insufficient. As such SVM was employed for the identification of those genera and species where spectrotyping was unsuccessful. Those genera and species are: *Achromobacter*, *Kingella*, *Salmonella*, *Serratia*, *Shigella*, *Stenotrophomonas*, *Burkholderia*, *Citrobacter*, *Enterobacter*, *Klebsiella*, *E.coli* and *P. aeruginosa*. Using a similar approach as the multitier pairwise spectrotyping method to create the spectral database, for the SVM models, multiple SVM models were created and coupled together to create a full spectral database. The kernel function for all SVM models used was Linear with a cost value of 1 using the spectral features of a broad region 1480-980 cm^{-1}. As a result of the model construction, 3 tiers (or levels) were the most optimal for genus bacterial identification application. Tier 1 is a single SVM model for the classification of multiple classes, in this case, the spectral groups were separated into those classes found in Tier 2; *Achromobacter*, *Kingella*, *Salmonella*, *Serratia*, *Shigella*, *Stenotrophomonas*, Group 1 and Group 2. Group 1 of spectra belongs to those belonging to selected *Enterobacteriaceae* bacteria (*Citrobacter*, *Enterobacter*, *Escherichia coli* and *Klebsiella*) and Group 2 consists of spectra belonging to *P. aeruginosa* and *Burkholderia*. As an output of Tier 1, the SVM model provides a [1] predicted class and [2] most likely class. As mentioned earlier, the predicted class is based on predictions between pairs (e.g., *Achromobacter* vs. *Kingella*; *Achromobacter* vs. *Salmonella*; *Salmonella* vs. *Kingella*) while Most Likely predicted is based on the probability cut-off value of >0.5 (a default value). In reference to Figure 4.16, Tier 1 is defined as the “global multiclass SVM model”. Tier 2 consists of 7 pairwise SVM models where each class (genus) is compared against the other genera. For example, SVM1 model consists of 2 classes, first class contains the spectra belonging to *Achromobacter* while the second class contains spectra belonging to the “Others” spectra in Tier 2. The “Others” spectra in SVM1 are therefore spectra belonging to *Kingella*, *Salmonella*, *Serratia*, *Shigella*, *Stenotrophomonas*, Group 1 and Group 2. The same logic applies to the other SVM models (Figure 4.16).

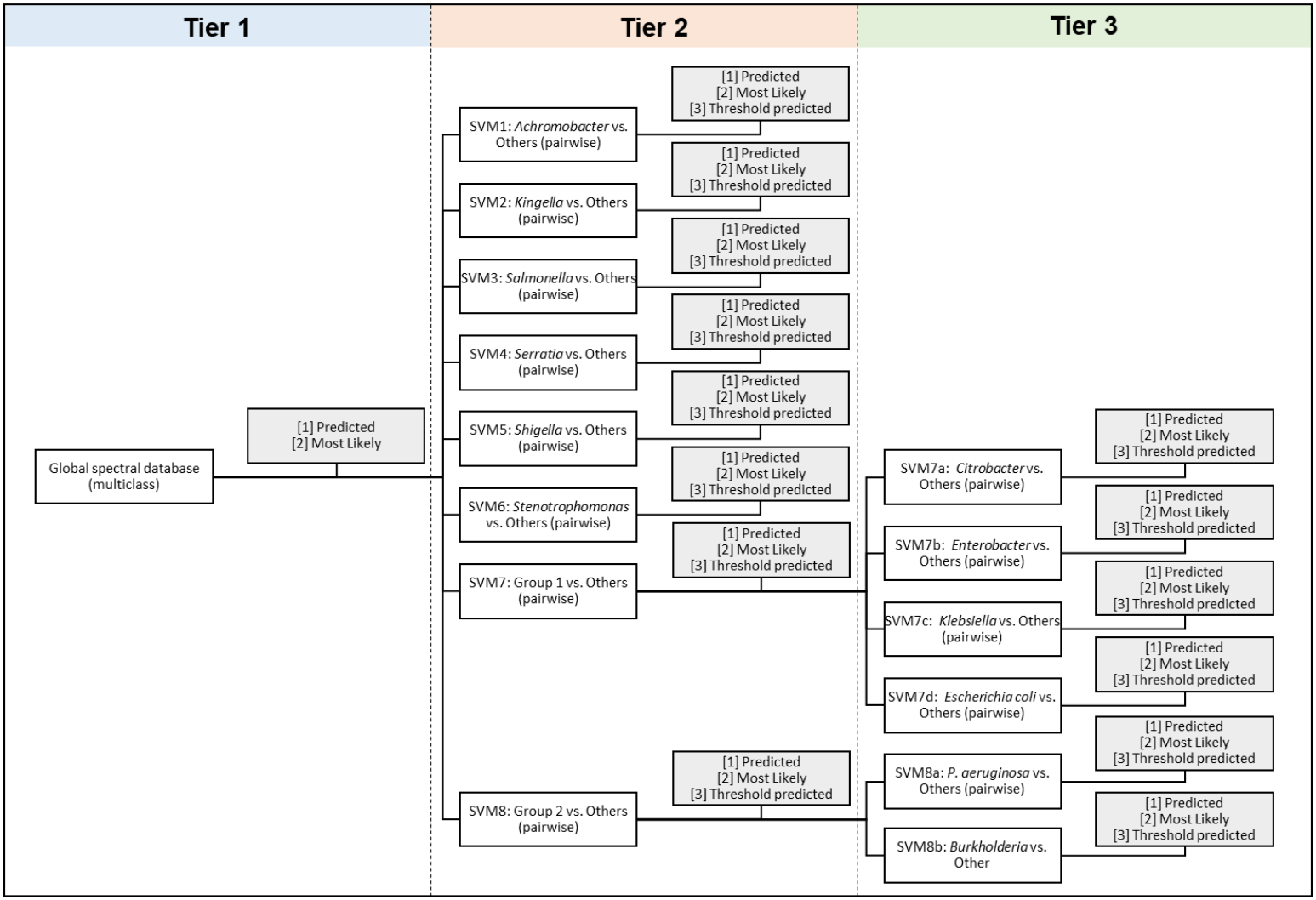


Figure 4.16. Global SVM spectral reference database structure for the classification of Set C2 Gram-negative bacteria where unknown spectra interrogate all SVM model(s) in Tiers 1 and 2 where SVM models are indicated in white rectangles and outputs are indicated by grey rectangles

Unlike multiclass SVM models, in addition to Predicted and Most Likely class output predictions, pairwise SVM models outputs a “threshold predicted” class, which is based on a threshold probability cutoff that is self-set (further details are found in CHAPTER 3).

Moreover, in Tier 2, SVM5 – *Shigella* vs. Others model consists of only two species, *S. sonnei* and *S. flexneri*, which by the spectrotyping technique, displays high discriminatory power (Figure 4.11). As such, once the unknown spectrum is predicted as *Shigella*, spectrotyping techniques are then employed as described for Set C1. Similarly, Tier 3 is also further broken down into Tier 4, for species identification of *Burkholderia* spp. (SVM8b-i) and *Klebsiella* spp. (SVM7c-i). For both SVM8b-i and SVM7c-i, the SVM models are multiclass to accommodate all species within the genus. Some species are not represented spectrally and therefore were omitted from the study along with those with limited number of spectra (SVM requires large data per class, at least triplicate spectra of 50 isolates per class).

For the creation of the SVM models, it should be noted for SVM7: Group 1 vs. Others, the training set had a 1.5% misclassification rate and a 9.3% misclassification rate during validation (Table 4.6). In addition, the validation set generalized R^2 is 0.57702 where closer to 1 is more significant. Spectra in Group 1 belongs to genera of the *Enterobacteriaceae* family; *Citrobacter*, *Enterobacter*, *Escherichia* and *Klebsiella*. It should be noted that these spectra were difficult to discriminate pairwise using the spectrotyping technique and SVM demonstrating the difficulty associated with classifying these microorganisms from the others in the SVM model. While SVM algorithms are more powerful than the feature selection algorithm used in the study, both methods suggest high spectral similarities between genera and species of the *Enterobacteriaceae* family and those that are Gram-negative. Compared to the other SVM models in Tier 2 (Figure 4.16) with less than 0.8% misclassification rates and a generalized R^2 values of >0.90 of the validation models, both SVM7 Group 1 and SVM8 Group 2 models were the least performing with 3.5-9.3% misclassification rates and a generalized R^2 values of <0.79 (Table 4.6). These results are not surprising as the genera within each of the groups are biochemically weak and are similar to the other GN bacteria and may be reflected in the ATR-FTIR spectra. As powerful as SVM algorithms are, it is limiting for those groups as

the data for differentiation are not unique enough in the spectra employed in training the models.

Moreover, within Group 1 for genus identification of those genera belonging to the *Enterobacteriaceae* family, the misclassification rates are approximately 6.0-9.0% with generalize R^2 values in the range of 0.50 to 0.86 (Table 4.6). With the removal of the other GN-negative bacteria (other genera that is not in Group 1), the SVM models did not improve, further agreeing the fact that with perhaps, not enough spectral information for high confidence discrimination between various GN bacteria genera. There is a chance where current biochemical identification methods have not identified the microorganisms correctly and the SVM models are training models on inaccurately identified samples. Further studies with absolute identification of the microorganisms will necessitate further conclusion of the spectral database for these 4 species.

To this end, a final ATR-FTIR spectral database was constructed with the use of combining both spectrotyping and SVM techniques. The second part of the study is to evaluate the performance of the developed ATR-FTIR spectra database on a subset of spectra that were not used in the spectral database but were coming from the same repository of spectra collected over the three years of the study.

Table 4.6. Global and pairwise SVM models (linear; cost=1) performances for genus identification

Tier 1: Global multiclass SVM model (training: 1031; validation: 258)				
SVM Model - Genus (support vectors; total no. isolates for validation)	Training misclassification	Validation misclassification	Validation Generalized R^2	Test misclassification (p=0.05)
Multiclass (#SV=339; *n*=258; test = 94)	0.0000	0.03488	0.98733	0.02128

Tier 2: Pairwise SVM models 1 through 8 (training: 1031; total validation: 258)				
SVM Model - Genus (support vectors; total no. isolates for validation)	Training misclassification	Validation misclassification	Validation Generalized R^2	(Probability) and threshold misclassification
SVM1 - *Achromobacter* (#SV=41; *n*=25)	0.00000	0.00390	0.94100	(0.8043) 0.00390
SVM2 - *Kingella* (#SV=18; *n*=7)	0.00000	0.00775	0.93503	(0.9022) 0.00775
SVM3 - *Salmonella* (#SV=59; *n*=42)	0.00000	0.00775	0.92221	(0.0922) 0.00301
SVM4 - *Serratia* (#SV=51; *n*=6)	0.00000	0.00775	0.89640	(0.6739) 0.00388
SVM5 - *Shigella* (#SV=44; *n*=18)	0.00000	0.00388	0.97005	(0.8043) 0.00388
SVM6 - *Stenotrophomonas* (#SV=47; *n*=24)	0.00000	0.00388	0.95618	(0.8261) 0.00388
SVM7 - Group 1[1] (#SV=140; *n*=74)	0.01455	0.09302	0.57702	(0.5761) 0.08915
SVM8 - Group 2[2] (#SV=87; *n*=62)	0.00000	0.03488	0.78748	(0.5109) 0.03101
Tier 3: Pairwise SVM models 7a through 7d (training: 315; validation: 70)				
SVM7a - *Citrobacter* (#SV=52; *n*=7)	0.00000	0.05714	0.50102	(0.7826) 0.04286
SVM7b - *Enterobacter* (#SV=54; *n*=12)	0.00000	0.08571	0.62723	(0.9022) 0.05714
SVM7c - *Escherichia* (#SV=56; *n*=30)	0.00000	0.05714	0.85870	(0.7283) 0.08710
SVM7d - *Klebsiella* (#SV=41; *n*=21)	0.00000	0.07143	0.80555	(0.8261) 0.07143
Tier 4: Pairwise SVM models 8a and 8b (training: 272; validation: 57)				
SVM8a - *Pseudomonas* (#SV=81; *n*=38)	0.00000	0.03509	0.95577	(0.8043) 0.00000
SVM8b - *Burkholderia* (#SV=81; *n*=19)	0.00000	0.03509	0.95577	(0.8043) 0.05300

[1] Group 1: *Enterobacteriaceae* consisting of *Citrobacter*, *Enterobacter*, *Escherichia* and *Klebsiella*

[2] Group 2: NFGNB consisting of *P. aeruginosa* and *Burkholderia* species

4.4.2. Evaluation of the constructed ATR-FTIR spectral database

Combining all spectral databases (Set A, Set B, Set C, SVM 1-7, SVM 7a-c and SVM 8a-b), an evaluation set of spectra was set aside for evaluating the spectral reference database as a whole. A total of 584 averaged spectra (averages of triplicate spectra collected from an isolate) were employed for the evaluation at the Gram-stain type level and achieved 99.3% correct identification with 0.17% misidentification and 0.52% no identification rates (Table 4.7). Breaking down the stats in reference to Table 4.7, the performance between GN and GP gram-type identification were similar, however, there were no misidentification observed for GP bacteria while GN bacteria had 1 spectrum from 1 isolate with a major misidentification (as GP bacteria). At the genus level, the misidentified spectrum is actually a *Pseudomonas* that is predicted as *Staphylococcus*; at the Gram-stain level, predicting with 98% confidence (based on the absolute cosine spectral similarity x 100) with a standard deviation from the mean of the reference GP spectra of -0.62169 (high confidence misidentification). According to the outlined confidence limits of CHAPTER 3, this result is conclusive and is a true misidentification. The 3 inconclusive results however, all predicted as the opposite Gram-stain type with low spectral similarity and are considered inconclusive.

Table 4.7. Performance evaluation of Gram-type from the developed ATR-FTIR spectral reference database

	No. of isolates				
Gram-stain type	Database	Collected	Correct (%)	Misidentified (%)	Inconclusive (%)
GN	1226	272	270 (99.3)	1 (0.37)	1 (0.37)
GP	1358	312	310 (99.4)	0 (0.00)	2 (0.64)
Total	2584	584	580 (99.3)	1 (0.17)	3 (0.52)

Upon further spectral investigation, spectral variance between the average GN spectrum and misidentified spectrum obtained a higher variance compared to the average GP spectrum and the misidentified spectrum. Although the misidentified spectrum is confidently wrong, further investigation of the isolate is needed as spectral variances between GN and GP bacteria are extremely different in the multiple regions due to major cell wall biochemical differences. Due to limited resources, the isolate was not further

investigated by Gram staining or re-identified and is reported as misidentified in the current thesis.

The combined spectrotyping and SVM models also achieved an overall of 94.3%, 2.25% and 3.45% rates for correct, misidentified and no identification results at the genus level respectively (Table 4.10). A total of 583 isolates were utilized for the evaluation where only 579 isolates were represented with spectra from at least 10 unique strains; *Micrococcus* spp. is therefore omitted from the evaluation as there are only spectra from 6 isolates. For genus identification of GN bacteria, the constructed spectral reference database achieved 92.3%, 2.94% and 4.78% correct, misidentified and no identification results at the genus level respectively (Table 4.10). GP bacteria performed significantly better at the genus level with 96.1% correct identification, 1.63% misidentification rates and 2.28% inconclusive (no identification results) (Table 4.10).

In all inconclusive/no identification results, the standard deviation from the predicted reference mean were above 3 and therefore the predictions were non-reportable. It is interesting to note that the no identification rates are higher than the misidentification rates which suggests the ability to identify non-represented microorganisms rather than to misidentifying those that are not represented or utilized for the construction of the spectral reference database. Moreover, in data not shown, the latter statement was confirmed by interrogating spectra belonging to species not represented in the spectral reference database, resulted in a combination of low spectral similarity and high standard deviation from the predicted reference mean spectrum.

Additionally, species diversity within a genus may also be attributed to the high misidentification rates of *Citrobacter*, *Klebsiella*, *Bacillus* and *Streptococcus* spp. These genera have multiple species represented in the spectral reference database but, lack representation of each species to further the identification at the species level. For example, there are 24 isolates of *Bacillus* in the spectral reference database, however, there are 4 species within the *Bacillus* genus represented. The 4 species are *Bacillus cereus* complex: *B. circulans*, *B. megaterium* and *B. thuringiensis* with 13, 1, 7 and 3 isolated represented respectively. Through spectral analysis, these 4 species have low spectral similarities and 3 of the 4 species would be considered underrepresented. The

same can be said for *Streptococcus* spp.; there are 20 group/species of *Streptococcus* where all but 2 species have over 10 isolates represented. If the cutoff of isolates represented in the spectral database is based on species within the genus, then the misidentification rates would be reduced.

Table 4.8. Genus-misidentified isolates and the predictions obtained from the ATR-FTIR spectroscopy-based microorganism identification technique and the confidences of the predictions

ID no.	Actual	Predicted	Confidence[1]
F_E002	*Citrobacter*	*Enterobacter*	65.0%
K_E002	*Citrobacter*	*Escherichia*	100%
O_E006	*Klebsiella*	*Enterobacter*	100%
P_E004	*Klebsiella*	*Escherichia*	75.0%
PA_E016	*Pseudomonas*	*Achromobacter*	100%
PA_E024	*Pseudomonas*	*Staphylococcus*[2]	98.0%
GD_E001	*Salmonella*	*Shigella*	70.0%
SM_E017	*Stenotrophomonas*	*Kingella*	70.0%
S_002	*Bacillus*	*Staphylococcus*	99.5%
E_047	*Enterococcus*	*Streptococcus*	99.5%
VG_002	*Streptococcus*	*Staphylococcus*	99.6%
VG_004	*Streptococcus*	*Staphylococcus*	99.7%
VG_006	*Streptococcus*	*Enterococcus*	99.3%

[1]Confidence based on the absolute cosine spectral similarity x 100
[2]Major error at Gram-stain level; high confidence misidentification based on high spectral similarity (confidence) and low standard deviation from the reference mean spectrum (-0.62169)

Upon closer inspection, a total of 13 isolates were misclassified, where 5 *Enterobacteriaceae* were misidentified as other genera within the *Enterobacteriaceae* family. Two isolates of NFGNB bacteria were misidentified as others within NFGNB, and 2 isolates were misidentified between closely related genera (*Enterococcus* and *Streptococcus*), while the remaining 4 isolates were misidentified without known relationship between the actual and predicted. The spectra of the 4 isolates were of good spectral quality and a third method of microorganism identification may be required to elucidate the reason for misidentification.

In reference to Figure 4.15 (database structure), employing the developed prediction model, the genus misidentified *Enterococcus* spp. isolate was erroneously misidentified between the pair of genera *Enterococcus* spp. and *Streptococcus* spp. The regions of discrimination selected through the feature selection algorithm for the two genera are

1090-1100,1146-1152,1271-1277,1340-1346,1349-1355 cm^{-1}. Through optimization, by restricting, but broadening the region to 1090-1355 cm^{-1}, the misidentified isolate was correctly identified as *Enterococcus* through HCA (Figure 4.17). As such, the current model does not make no mistakes; optimization of selected regions may be further investigated to improve the performance of the prediction model. Although the current model did not correctly identify the genus identification of the microorganism, through spectral analysis, the information to discriminate between the two genera is available and can successfully differentiated spectra belonging to *Enterococcus* spp. and *Streptococcus* spp.

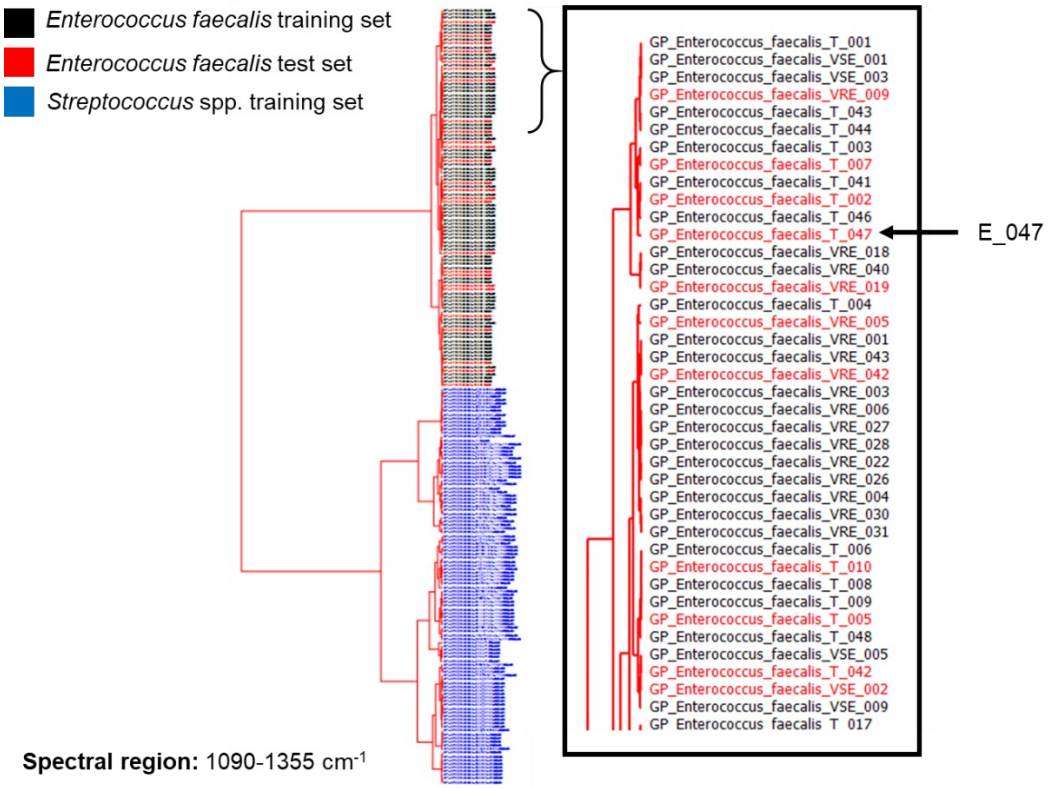


Figure 4.17. Further investigation of genus misidentified *Enterococcus* spp. isolate E_047 using a restricted broad region demonstrating correct genus clustering between predicted genus (*Streptococcus*) and actual genus

Moving forward to evaluating the constructed ATR-FTIR spectral database for species identification, only species with spectra represented with over 10 unique isolates were used for the evaluation. After filtration, a total of 14 genera (10 GN bacteria and 4 GP bacteria) and 25 species (14 GN bacteria and 11 GP bacteria) were remaining from the

previous spectral dataset (Table 4.11). In terms of isolates, there are a total of 1421 isolates represented of which, 699 belonging to GN bacteria and 722 from GP bacteria. The overall performance rates for correct, miss and no identification are 95.9%, 1.37% and 2.75% respectively (Table 4.11). Once again, species identification of GP bacteria using the spectral reference database is performing significantly higher than GN bacteria by 6.5%.

Table 4.9. Species-misidentified isolates and the predictions obtained from the ATR-FTIR spectroscopy-based microorganism identification technique and the confidences of the predictions

ID no.	Actual	Predicted	Confidence	Misidentification Class
KO_E006	*K. oxytoca*	*Enterobacter species*	100%	Class V
KP_E004	*K. pneumoniae*	*E. coli*	77%	Class V
KP_E005	*K. pneumoniae*	*E. coli*	60%	Class V
PA_E024	*P. aeruginosa*	*S. aureus*	98%	Class IV
SGD_E001	*Salmonella enterica*	*Shigella species*	70%	Class IV
SM_E017	*S. maltophilia*	*K. kingae*	70%	Class IV

Of the 6 misidentified isolates, similar to the genus misidentifications, those species belonging to the *Enterobacteriaceae* family were being misidentified as other species within the same family, the same goes for those bacteria NFGNB, with the exception of isolate PA_E024 (Table 4.9). Isolate PA_E024 was also incorrect at the genus level. Misidentification/classifications are further detailed in Table 4.12. Increasing the spectral representation of species such as *Citrobacter freundii*, species of *Klebsiella* and *P. aeruginosa* may facilitate higher performance of the database for GN species identification. *P. aeruginosa* is an extremely diverse species and representation of the species was not examined in this thesis. Other methods of species' identification may be required to ascertain the identity of the microorganisms such as with the use of whole genome sequencing.

Overall, the construction of the ATR-FTIR spectral database was evaluated and achieved 95% correct genus identification and 94.3% correct species identification. As such, the ATR-FTIR spectral reference database was further evaluated in a clinical setting over a 3-month period to determine the robustness of the standardized technique and method.

Table 4.10. Evaluation table for the performance of the combined spectrotyping and SVM models for the ATR-FTIR spectroscopy-based method for genus identification

Gram-stain type	Genus	No. of isolates (%)				
		Database	Collected	Correct ID	Misidentification	No identification
GN	*Achromobacter*	72	25	25	0	0
	Acinetobacter	14	5	5	0	0
	Burkholderia	100	24	23 (95.8)	0	1 (4.17)
	Citrobacter	19	8	4 (50)	2 (25)	2 (25)
	Enterobacter	54	13	11 (84.6)	0	2 (15. 4)
	Escherichia	164	31	30 (96. 8)	0	1 (3.23)
	Kingella	14	7	7	0	0
	Klebsiella	77	22	17 (77. 3)	2 (9.09)	3 (13.6)
	Pseudomonas	292	47	44 (93.6)	2 (4.26)	1 (2.17)
	Salmonella	145	42	39 (92. 9)	1 (2.38)	2 (4.76)
	Serratia	25	6	5 (83.3)	0	1 (16.7)
	Shigella	59	18	18	0	0
	Stenotrophomonas	118	24	23 (95.8)	1 (4.17)	0
GP	*Bacillus*	24	15	13 (86. 7)	1 (6.67)	1 (6.67)
	Corynebacterium	10	6	6	0	0
	Enterococcus	171	50	49 (98)	1 (2.00)	0
	Listeria	68	20	20	0	0
	Micrococcus	6	4	4	0	0
	Staphylococcus	502	164	164	0	0
	Streptococcus	102	52	44 (84.6)	3 (5.77)	5 (9.62)
Total	Total isolates	2036	583	551 (94.5)	13 (2.23)	19 (3.26)
	Total isolates represented	2030	579	547 (94.5)	13 (2.25)	19 (3.28)
GN	Total isolates represented	1153	272	251 (92.3)	8 (2.94)	13 (4.78)
GP	Total isolates	883	311	300 (96.5)	5 (1.61)	6 (1.93)
	Total isolates represented	877	307	296 (96.4)	5 (1.63)	6 (1.95)

Table 4.11. Evaluation table for the performance of the combined spectrotyping and SVM models for the ATR-FTIR spectroscopy-based method for species identification

Gram-stain type	Genus	No. of isolates (%)				
		Database	Collected	Correct	Misidentification	No identification
GN	*Burkholderia gladioli*	26	5	5	0	0
	Burkholderia multivorans	27	6	6	0	0
	Burkholderia species cepacia-complex	19	4	4	0	0
	Citrobacter freundii	13	6	4 (66.7)	0	2 (33.3)
	Escherichia coli	141	31	31	0	0
	Kingella kingae	14	7	7	0	0
	Klebsiella oxytoca	18	6	4 (66.7)	1 (16.7)	1 (16.7)
	Klebsiella pneumoniae	52	13	8 (61.5)	2 (15.4)	3 (23.1)
	Pseudomonas aeruginosa	42	47	44 (93.6)	1 (2.13)	2 (4.26)
	Salmonella enterica	145	42	40 (95.2)	1 (2.38)	1 (2.38)
	Serratia marcescens	25	6	5 (83.3)	0	1 (16.7)
	Shigella flexneri	25	6	6	0	0
	Shigella sonnei	34	12	12	0	0
	Stenotrophomonas maltophilia	118	24	23 (95.8)	1 (4.17)	0
GP	*Bacillus cereus*	13	8	8	0	0
	Enterococcus faecalis	77	26	24 (92.3)	0	2 (7.69)
	Enterococcus faecium	94	22	22	0	0
	Listeria monocytogenes	57	13	13	0	0
	Staphylococcus aureus	333	110	110	0	0
	Staphylococcus capitis	20	5	5	0	0
	Staphylococcus epidermidis	44	13	13	0	0
	Staphylococcus haemolyticus	16	6	6	0	0
	Staphylococcus hominis	29	8	8	0	0
	Staphylococcus lugdunensis	24	6	6	0	0
	Staphylococcus warneri	15	5	5	0	0
TOTAL	Total isolates represented (%)	1421	437	419 (95.9)	6 (1.37)	12 (2.75)
GN	Total isolates represented (%)	699	215	199 (92.6)	6 (2.79)	10 (4.65)
GP	Total isolates represented (%)	722	222	220 (99.1)	0	2 (0.90)

Table 4.12. Classification and description of misidentification results

Misidentification class	
Class	Description
Class I[1]	True misidentification: routine and gold standard microbial identification methods have been executed and reanalysis of the isolates with routine and ATR-FTIR spectroscopy are completed and resulting similar results to initial results. In addition, results by spectral analysis of the misclassified bacteria by HCA must agree with the results from spectral database search in comparison to the averages of genus/species spectra in the spectral database. Isolates in Class 1 are eligible for additional experiments to ascertain biochemical differences (as observed in the ATR-FTIR spectra) from other strains used for the creation of the ATR-FTIR spectral database.
Class II[1]	Database misidentification: routine and gold standard microbial identification are in agreeance, reanalysis by routine and the ATR-FTIR microbial based method results are similar to those of initial analysis. However, upon further spectral analysis by HCA and PCA between those spectra in the spectral database and misidentified spectra, results coincide with routine and gold standard microbial identification methods. Isolates in Class II indicated the need for ATR-FTIR spectral reference database optimization by (i) increasing spectral representation of the pairs where misidentification is occurring (ii) restructuring/reselecting pairs of the multitier spectral database and/or (iii) improving feature selection between pairs where misidentification is occurring.
Class III[1]	Spectral similarity misidentification; known limited spectral dissimilarities between actual and predicted species. Class III misidentification indicates limitation of the outlined ATR-FTIR spectroscopy-based microbial identification method. Optimization of the spectral reference database may be explored to improve (or not) the prediction of the misidentified, however, external methods not discussed in the thesis may have to be explored for further work.
Class VI	Inconclusive misidentification: routine and ATR-FTIR spectroscopy-based microorganism identification is not in concordance, however, no tie breaker (gold standards) and/or reanalysis by standard routine methods were completed. Class VI misidentification are misidentifications that were not resolved and does not fully explain the inaccuracy of the method as conventional methods are known to also not produce 100% correct species identification.
Class V	Spectral similarity and inconclusive misidentification; a combination of both Class III and Class VI misidentification where there are known high spectral similarities however, no tie-breaker identification method was carried out to conclude the results.

[1]Class I, II and III are not presented in the study due to limitations: gold standard identification methods, access to laboratories and samples were no longer available during the course of the study.

4.4.3. 3-month clinical routine prospective evaluation of the constructed ATR-FTIR spectral reference database

The performance of the constructed database provided a 98.4% correct species identification for various common and clinically relevant GN and GP bacteria. Over a course of 3 months, a spectrometer was placed at Centre hospitalier Sainte Justine (CHUSJ) where ATR-FTIR spectra were collected on all routine isolates that were identified by VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile), VITEK® MS (bioMérieux, Marcy-l'Étoile) and/or through gene sequencing of targeted genes (e.g., *recA* for *Burkholderia* species). A total of 391 isolates were collected where 363 were represented in the spectral reference database. Overall, the prospective study resulted in 97.2% and 98.4% correct identification; 1.38% and 0.32% misidentification for genus and species respectively and a 1.34% no identification rate.

In routine, the spectral reference database performed greater than the evaluation especially for those genera and/or species within the *Enterobacteriaceae* family. Of the 363 isolates represented, 125 were GN bacteria with the remaining 238 identified as GP bacteria. For the GN bacteria, the prospective study achieved 100%, 95.2% and 97.5% correct Gram-stain type, genus, and species identification respectively with 2.40% misidentification rate at the genus level and 1.23% at the species level, and 2.40% no identification rate. While GP bacteria achieved 98.7%, 98.3%, 98.7% correct Gram-stain type, genus and species identification respectively; 0.42% and 0.84% misidentification rates at the Gram-stain type and genus respectively (no misidentification at the species level), and a 0.84% no identification rate.

In comparison to the evaluation set, the prospective set performed poorly for genera of species belonging to the *Enterobacteriaceae* family, in particular, *Citrobacter* and *Klebsiella.* Misidentification only occurred for *Acinetobacter* species, *K. oxytoca* and *S. sonnei*. It should be noted that both isolates that were misidentified belong to the *Enterobacteriaceae* family and were misidentified as other species within the same family. Similar to the evaluation set, all misidentified *Klebsiella* were identified as *E. coli*. Although belonging to the same family, *E. coli* and *Klebsiella* phenotypically differ in motility

(*Klebsiella* being non-motile while *E. coli* is motile) (73). Spectrally, these two spectral groups are not the closest neighbor from one another (Figure 4.10).

Table 4.13. 3-month prospective study species-misidentified isolates and the predictions obtained from the ATR-FTIR spectroscopy-based microorganism identification technique and the confidences of the predictions

Sample ID	Actual	Predicted	Confidence	Misidentification class
ABC_P001	*Acinetobacter* species	*Enterobacter* species	85%	Class IV
KO_P003	*K. oxytoca*	*E. coli*	65%	Class V
SS_P003	*S. sonnei*	*E. coli*	100%	Class V
BS_P001	*Bacillus* species	*Corynebacterium* species	100%	Class IV
SA_P121	*S. aureus*	*Achromobacter* species	100%	Class IV

Moreover, *E. coli* and *Shigella* differ phenotypically in many aspects though they are phylogenetically similar where *Shigella* is said to have diverged from *E. coli*. *Shigella* is non-capsulated, motile and does not ferment lactose (73). Conventionally, for differentiation between the two microorganisms, one would look for whether the isolate contains lysine decarboxylase and lactase, motility and/or utilizes L-serine, D-xylose and sodium acetates as nutrient sources (74). Other studies have successfully discriminated between the two by PCR gene amplification of the *uidA* gene (encodes for B-glucuronidase) and the *lacY* (encodes for lactose permease) gene (75). As observed in Figure A.6, the spectral dissimilarities between the two species are significant for species differentiation. *E. coli* and *Shigella* are phylogenetically similar (from 80-90% nucleotide similarity (76)), if not, *Shigella* to some microbiologists are variants of *E. coli*, however, in clinical practice, *Shigella* are biochemically dissimilar to *E. coli* (73).

MALDI-TOF MS, which identifies microorganisms based on protein mass fingerprints has revolutionized clinical microbiology with its rapid microorganism identification technique, however, the widely used technology has the inability to accurately discriminate between *E. coli* and *Shigella* (77). Unsuccessful discrimination between the two microorganisms by MALDI-TOF MS may be due to the fact that they share similar protein profiles and also cannot be discriminated by 16S rRNA gene sequencing (78). Recent findings have however, differentiated between the two species with whole-genome sequencing,

suggesting that a protein-specific method is not sufficient for species differentiation while whole organism analysis is more reliable (65).

In summary, the prospective study shows that 1 isolate of *S. sonnei* (isolate SS_P003) collected in routine was incorrectly identified as *E. coli* (Table 4.13). Upon further spectral analysis, utilizing the region for discriminating between *E. coli* and *Shigella* (997-1003, 1025-1036, 1165-1174, 1178-1187, 1467-1472 cm^{-1}) and the average spectra of *E. coli* and *Shigella* from the spectral database achieved 99.9% spectral similarity with the nearest neighbor *E. coli* with the misidentified *S. sonnei*. However, at a broader region (which still discriminates between *E. coli* and *Shigella* of the spectral reference database), 1380-980 cm^{-1}, there is a reduction in spectral similarity between *E. coli* and the misidentified SS_P003 isolate of 62% spectral similarity and 51% spectral similarity to *Shigella* genus. Moreover, using the broad region, the spectral similarity between *E. coli* and non-O157 *E. coli* is 98%. As such, it is possible that based on the spectra accumulated for both species (database and evaluation sets), the observed misidentified *S. sonnei*, is either a unique strain that is not represented in the spectral database that is closer to *E. coli* or the isolate is a mixed culture displaying identical colony morphology on the culture media plate. Based on the spectral analysis, an additional method of species identification is needed to ascertain the discrepancy and further phenotypic analysis (antimicrobial susceptibility testing (AST), biochemical tests, cell morphology, serotyping, etc.). The misidentified isolates were not reanalyzed by gold standard methods to resolve the discrepancies or to settle the predictions between conventional methods and the ATR-FTIR spectroscopy-based method.

Additionally, of the 216 collected routine *S. aureus*, 1 isolate was misidentified at the gram-level (therefore genus and species levels as well) and 1 isolate provided no identification (Table 4.13). The misidentified SA_P121 isolate was identified as a Gram-negative *Achromobacter* species (this genus does not have identification output past the genus level). Upon further spectral investigation, sample SA_P121's closest neighbor by HCA at the Gram-stain level is Gram-negative bacteria with 98% spectral similarity using a broad region of 1380-980 cm^{-1} ((A) of Figure 4.18). Using only the average spectra of the reference Gram-negative bacteria and the spectrum SA_P121, sample SA_P121's

closest neighbor with ~98% spectral similarity is *Achromobacter* species utilizing a broad spectral region ((B) of Figure 4.18).

Through spectrotyping, SA_P121 was misidentified in Set A (Gram-stain type) of the spectral database and water fell from there (Table 4.13). Further investigation of the isolate in an SVM pairwise model for the classification of Gram-stain type (Gram-negative and Gram-positive bacteria; omitting Gram-stain variable bacteria), resulted in the same misidentification.

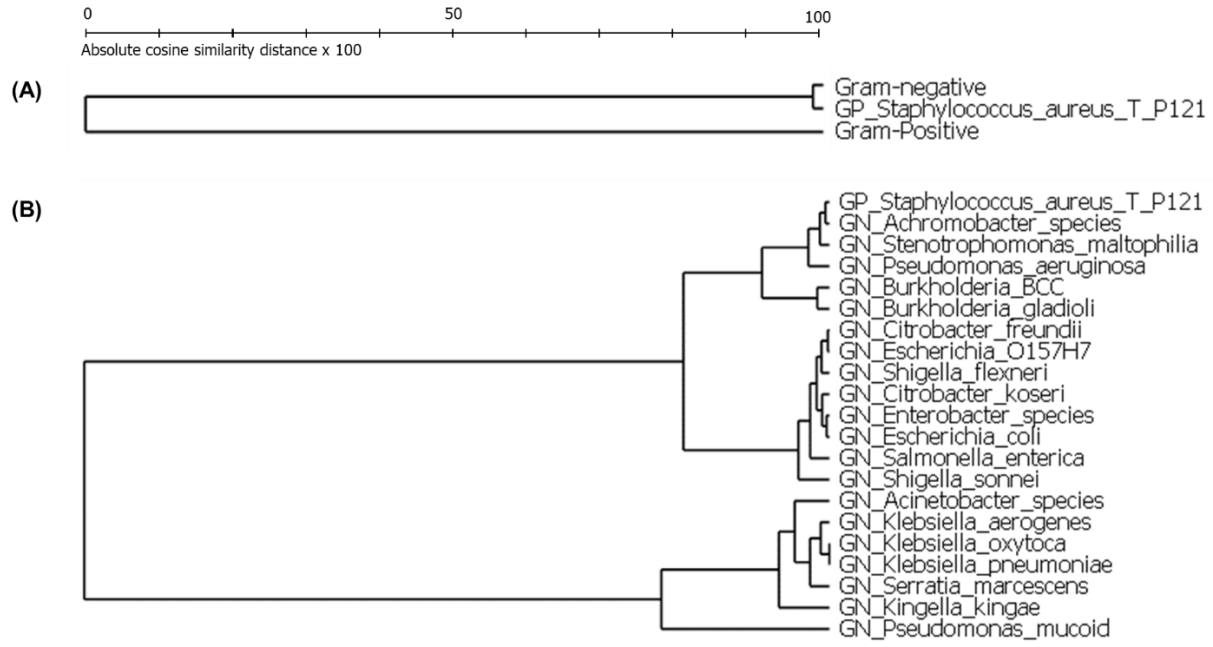


Figure 4.18. Investigation of SVM model misidentification of routine identified *Staphylococcus aureus* (SA_P121) through spectrotyping using the region 1380-980 cm^{-1}. Dendrogram represents HCA of (A) SA_P121 clustering closest to average reference spectrum of Gram-negative bacteria and (B) SA_P121 clustering closest with average reference spectrum of Gram-negative bacteria species

Through spectrotyping and being identified as Gram-negative bacteria, the SA_P121 spectrum then went through the interrogation of the SVM models described in section 4.4.1.4 – resulting in the misidentification of the isolate as *Achromobacter*. Using the same method of spectral investigation through spectrotyping of the SVM-misidentified sample, the misidentified *S. sonnei* (isolate SS_P003) also resulted in similar results, therefore the misidentification by the SVM model was also observed for the spectrotyping technique and vice versa.

Furthermore, 1 isolate of routine identified *S. aureus* (SA_P073) resulted in no identification by spectrotyping as displayed in Figure 4.19-(A), no spectral similarities were observed between the averages of Gram-positive and Gram-negative bacteria through HCA. In Figure 4.19-(B), the HCA between those Gram-positive bacteria and the SA_P073 spectrum also indicates no spectral similarities with the Gram-positive species represented in the spectra database.

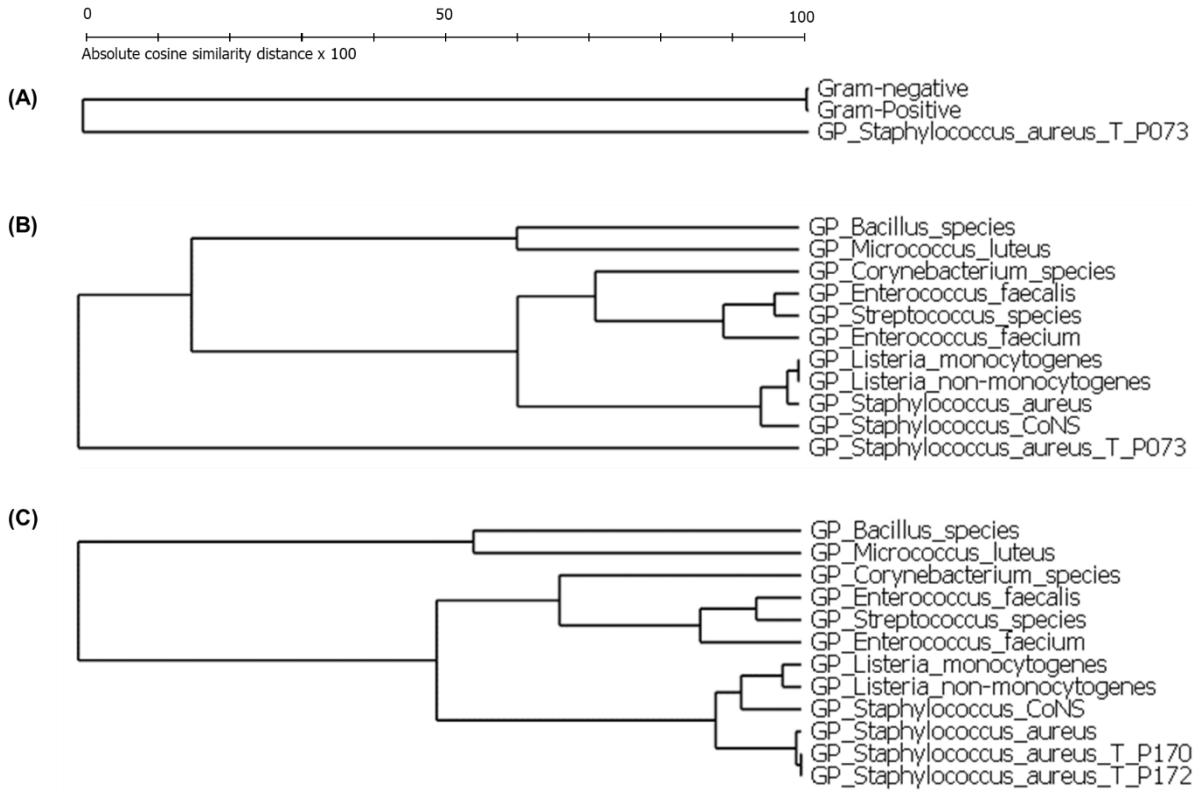


Figure 4.19. Investigation of spectrotyping-based spectral database of non-identified spectrum, of routine identified *Staphylococcus aureus* (SA_P073) through spectrotyping using the region 1380-980 cm^{-1}. Dendrogram represents HCA of (A) SA_P073 clustering with neither of the average reference spectrum of Gram-negative bacteria and Gram-positive bacteria, (B) SA_P073 clustering with none of the average reference spectrum of Gram-positive bacteria species and (C) illustrates ATR-FTIR spectroscopy-based microorganism in agreeance with routine identified *S. aureus* (P170 and P171) clustering closest with *S. aureus* in the spectral database

While (C) of Figure 4.19 illustrates high spectral similarity between the 2 (isolates P170 and P171) correctly identified prospective isolates (relative to routine identification results) using the spectrotyping method. Moreover, average spectra of yeasts and molds were introduced to the Gram-positive and Gram-negative averages and sample SA_P073

resulted in ~80% spectral similarity between the fungi versus bacteria (Figure 4.20). The lack of an identification result for SA_P073 isolate is based on the low cosine similarity distances and a high standard deviation from the reference predicted spectrum mean as described in detail in CHAPTER 3.

Moreover, all non-represented species were correct at the Gram-stain level and no-identification is defined as not being identified as anything that is represented in the spectral reference database. As such, the outlined method seldom (relative to the evaluation set) misidentifies non-represented microorganisms. Unfortunately for both the erroneous misidentification and no identification results of the study, no further analysis was carried out to ascertain the discrepant results between the two methods of identification (routine and ATR-FTIR spectroscopy-based).

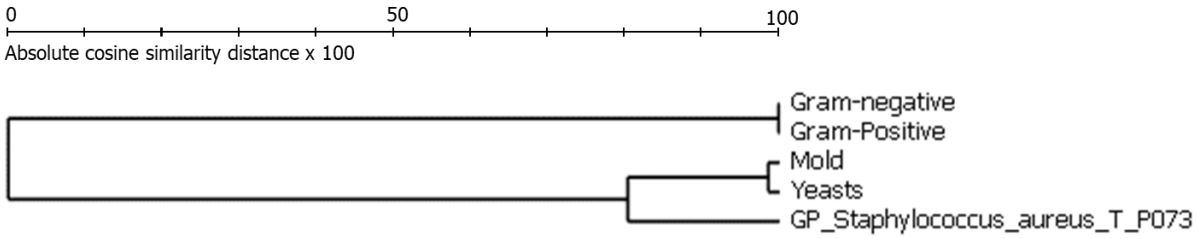


Figure 4.20. Dendrogram generated from hierarchical cluster analysis for the investigation of spectrotyping-based spectral database of non-identified spectrum of routine identified *Staphylococcus aureus* (SA_P073) through spectrotyping using the region 1380-980 cm^{-1} of average groups of reference spectra

Although some results by the ATR-FTIR spectroscopy-based method of microorganism identification were originally discordant with reference methods, upon further investigation, identification by ATR-FTIR spectroscopy was proven to be correct. For example, Figure 4.21 illustrates isolate NS115 provided by one of the participating institutions and labelled as *E. faecium*, however, through spectrotyping, it was closely clustering with *S. aureus* from the same set of samples. Upon further investigation, the institution made a human-error of mislabelling the identification of isolate NS115, when in fact, it is a *S. aureus* isolate. Several similar instances were observed through the course of the thesis preparation and isolates that were mislabelled by the institution were omitted. Spectrotyping in this case offers rapid screening of various microorganisms and can flag

samples to be further investigated. Human-error may be due to mislabelling and/or placing the wrong sample in the storage tube, as such, reporting these errors are not part of the thesis, however, it does highlight the application of the technique and was used when filtering the data.

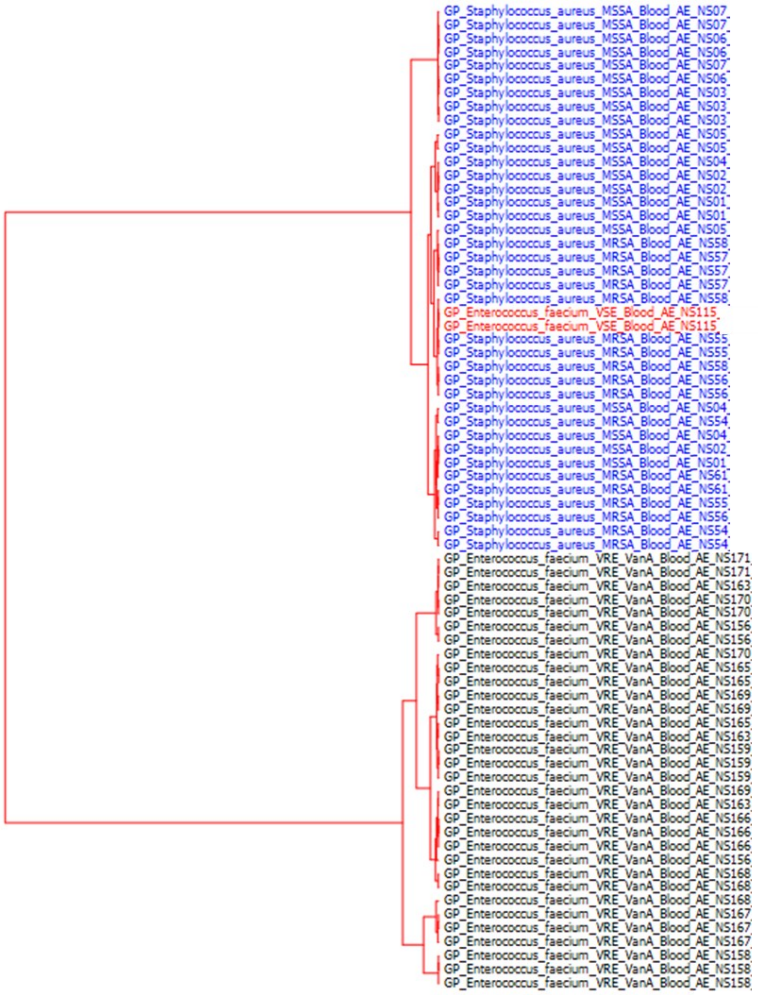


Figure 4.21. Example of participating institution-mislabeled sample N115 as *Enterococcus faecium* and correctly identified (and validated) as *Staphylococcus aureus* by the ATR-FTIR spectroscopy-based microorganism identification technique

Although *Micrococcus* species are underrepresented in the spectral reference database (with spectral representation from 6 isolates), the one isolate that was collected in routine was correctly identified. As such, this suggests the potential need for species-specific reference database cut-offs in terms of how many isolates are needed to be considered representative for each species. As observed in the evaluation set, there are spectra

belonging to 52 isolates of *Klebsiella pneumoniae*, however, there are 2 isolates from the evaluation set that were misidentified. Species within the genera *Klebsiella* exhibit higher intra species diversity and may need more than 52 isolates (of varying diversity) to be representative. Conversely, the current 52 isolates may not be diverse in phenotype and if isolates of *K. pneumoniae* were carefully selected to represent the common phenotypes, it may only need fewer isolates for complete spectral representation of the species.

Additionally, average spectra were utilized for the prospective study, however, non-averages were also evaluated and achieved identical results (with the exception of confidence values, however, they are within range) with a slightly different approach. As there are triplicate spectra per sample, no identification is attributed to triplicate spectra all having different identifications and/or low cosine similarity distance and high standard deviation from predicted mean reference spectrum of 2 or more spectra. The confidence (absolute cosine similarity distance x 100) are the averages of the replicates that fit the no-identification criteria.

Table 4.14. 3-month clinical routine prospective evaluation table for the performance of the combined spectrotyping and SVM models for the ATR-FTIR spectroscopy-based method for bacterial identification

		No. of isolates (%)								
				Correct (%)			Misidentification			No ID
Gram-stain type	Microorganism	Database	Collected	Gram	Genus	Species[4]	Gram	Genus	Species[4]	
GN	*Acinetobacter species*[1]	14	4	4	3 (75)	-	0	1 (25)	-	0
	Bacteroides fragilis	0	2	2	0	-	0	0	-	2
	Burkholderia species	102	1	1	1	-	0	0	-	0
	Citrobacter farmeri	0	1	1	1	-	0	0	-	0
	Citrobacter freundii	13	3	3	1 (33.3)	-	0	0	-	2 (66.7)
	Citrobacter youngae	0	3	3	0	-	0	0	-	3
	Enterobacter species[2]	54	25	25	25	-	0	0	-	0
	Escherichia coli[3]	24	9	9	9	9	0	0	0	0
	Escherichia coli	141	26	26	26	26	0	0	0	0
	Klebsiella oxytoca	18	8	8	7 (87.5)	-	0	1 (12.5)	-	0
	Morganella morganii	0	7	7	0	-	0	0	-	7
	Pantoae species	0	4	4	0	-	0	0	-	4
	Paracoccus yeei	0	2	0	0	-	0	0	-	2
	Proteus hauseri	0	1	1	0	-	0	0	-	1
	Proteus mirabilis	0	2	2	0	-	0	0	-	2
	Salmonella enterica	145	10	10	10	10	0	0	0	0
	Serratia marcescens	25	2	2	2	2	0	0	0	0
	Shigella flexneri	25	9	9	9	9	0	0	0	0
	Shigella sonnei	25	7	7	6 (85.7)	6 (85.7)	0	1 (14.3)	1 (14.3)	0
	Shigella species	50	3	3	3	-	0	0	-	0
	Stenotrophomonas maltophilia	118	18	18	17 (94.4)	17 (94.4)	0	0	0	1 (5.56)
GP	*Bacillus species*	24	1	1	0	-	0	1	-	0
	Enterococcus faecalis	77	19	19	19	19	0	0	0	0
	Enterococcus faecium	47	1	0	0	0	0	0	0	1
	Lactococcus species	0	1	1	0	-	0	0	-	1
	Micrococcus luteus	6	1	1	1	-	0	0	-	0
	Micrococcus lylae	0	2	2	2	-	0	0	-	0
	Rhodococcus equi	0	1	1	0	-	0	0	-	1
	Rothia mucilaginosa	0	1	1	0	-	0	0	-	1
	Staphylococcus aureus	334	216	214 (99.1)	214 (99.1)	214 (99.1)	1 (0.46)	1 (0.46)	0	1 (0.46)
	Streptococcus mitis	11	1	1	1	-	0	0	-	0
Total	Total isolates (%)	1253	391	386 (98.7)	357 (91.3)	312 (98.4)[5]	1 (0.26)	5 (1.28)	1 (0.32)[5]	29 (7.42)
	Total isolates represented (%)	1247	363	360 (99.2)	353 (97.2)	312 (98.4)[5]	1 (0.28)	5 (1.38)	1 (0.32)[5]	5 (1.34)
GN	Total isolates (%)	754	147	145	120 (82.8)	79 (97.5)[6]	0 (0.00)	3 (2.07)	1 (1.23)[6]	24 (16.6)
	Total isolates represented (%)	754	125	125	119 (95.2)	79 (97.5)[6]	0 (0.00)	3 (2.40)	1 (1.23)[6]	3 (2.40)
GP	Total isolates (%)	499	244	241 (98.0)	237 (96.3)	233 (94.7)[7]	1 (0.41)	2 (0.81)	0 (0.00)[7]	5 (2.03)
	Total isolates represented (%)	493	238	235 (98.7)	234 (98.3)	233 (98.7)[7]	1 (0.42)	2 (0.84)	0 (0.00)[7]	2 (0.84)

Note: Footnotes associated with the table are located on the following page

[1]Species all belongs to the *Acinetobacter calcoaceticus-baumannii* complex

[2]Species belonging to the *Enterobacter cloacae*-complex

[3]*E. coli* identified to the serotype level as *E. coli* O157:H7

[4]Not all genera are predicted to the species level, those that are not predicted to the species level are indicated with a hyphen in the cell of the table

[5]A total of 317 isolates were represented at the species level

[6]A total of 81 isolates were represented at the species level for Gram-negative bacteria

[7]A total of 236 isolates were represented at the species level for Gram-positive bacteria

4.5. CONCLUSION

The current classification of microorganisms (classical taxonomy) greatly aids in the selection of groups of spectra to form pairs for the pairwise multitier spectral database structure. Although current classification of microorganisms aids the database structure, there are some exceptions where it is not useful such as for those species with great biochemical diversity (*P. aeruginosa*) or other species displaying extremely similar spectral fingerprints (genera within the *Enterobacteriaceae* family) where powerful methods such as SVM models are employed to assist in species identification. As such, using the pairwise multitier method, the database construction is predominantly based on spectrotyping and machine learning. It should be interesting to note that SVM may be used to confirm spectrotyping-based results or even be used to construct the entire spectral reference database, however, spectrotyping offers additional spectral information and rapid screening capabilities based on nothing more than the comparison of the absorbance bands and relative intensities between groups of spectra.

Based on the results of the study, there is a high correlation between taxonomy classification and spectrotyping such that current taxa is based on phenotypic characteristics and ATR-FTIR spectra of microorganisms are associated with biochemical diversity which can be considered to be a phenotypic method. However, species classification is fluid and is ever changing with increased interest in classifying microorganisms based on genetic relatedness; while in the current study, classification of microorganisms was based on their spectral profiles. ATR-FTIR spectrotyping may have opened doors to applying such method for strain-typing microorganisms with spectral region selection and be used as a classification, pre-screening and strain-typing tool.

Although predictions at both a retrospective and prospective evaluation of the spectral reference database were >90% correct, there are limitations for various GN bacteria species belonging to the *Enterobacteriaceae* family. For future developments, isolates collected to construct the spectral reference database will require to have the complete pedigree of isolates such as the AST profiles, geographical origin of isolation, strain-type, serotype, source of isolation, biochemical tests profiles and others. In addition to requiring the listed (and more) information, the isolates would have to be carefully selected when

creating a spectral database to limit redundant representation. This will allow for the development of species-specific representation cut-offs, or how many isolates are needed of a specific species to represent the species in complete. The additional information may also become useful for datamining and exploring ATR-FTIR spectra for typing of various features such as the AST profiles to discriminate between antimicrobial susceptible and resistant microorganisms. It may also be possible to extenuate difference between closely related species by growth of these microorganisms on selective media.

With the limitation of being unable to reanalyze the misidentified and no identification isolates by conventional, gold standard and the ATR-FTIR spectroscopy-based method, an overall 97.2% correct genus identification, 98.4% species identification and no misidentification of non-represented microorganisms in routine, the study may be considered a success and may potentially be further explored for routine use. It is a rapid (1-minute analysis after initial culture) which is reagent-free, easy-to-use, specific, easy-to-train, compact and is cost effective compared to currently available non-spectroscopy-based methods. Although it is incomplete for species identification on a routine basis with the lack of representation of various species, it may be useful for screening and discriminating between difficult to differentiate *E. coli* and *Shigella*, and for serotyping while building a larger database.

The technology and technique will greatly benefit resource limited microbiology laboratory without compromising the specificity of the identification of microorganisms and can be used side-by-side with current identification techniques to confirm species identification.

4.6. REFERENCES

1. Kramer A, Schwebke I, Kampf G. 2006. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. BMC Infect Dis 6:130.
2. Juyal D, Prakash R, Shanakarnarayan SA, Sharma M, Negi V, Sharma N. 2013. Prevalence of non-fermenting gram negative bacilli and their in vitro susceptibility pattern in a tertiary care hospital of Uttarakhand: A study from foothills of Himalayas. Prevalence 2:108-112.
3. Rocchetti TT, Silbert S, Gostnell A, Kubasek C, Jerris R, Vong J, Widen R. 2018. Rapid detection of four non-fermenting Gram-negative bacteria directly from cystic fibrosis patient's respiratory samples on the BD MAX™ system. Practical laboratory medicine 12:e00102.
4. Gautam V, Sharma M, Singhal L, Kumar S, Kaur P, Tiwari R, Ray P. 2017. MALDI-TOF mass spectrometry: An emerging tool for unequivocal identification of non-fermenting Gram-negative bacilli. The Indian Journal of Medical Research 145:665.
5. Fernández-Olmos A, García-Castillo M, Morosini M-I, Lamas A, Máiz L, Cantón R. 2012. MALDI-TOF MS improves routine identification of non-fermenting Gram negative isolates from cystic fibrosis patients. Journal of Cystic Fibrosis 11:59-62.
6. Cutting GR. 2015. Cystic fibrosis genetics: from molecular understanding to clinical application. Nature Reviews Genetics 16:45-56.
7. Cohn JA. 2005. Reduced CFTR function and the pathobiology of idiopathic pancreatitis. Journal of clinical gastroenterology 39:S70-S77.
8. Cassagne C, Cella AL, Suchon P, Normand AC, Ranque S, Piarroux R. 2013. Evaluation of four pretreatment procedures for MALDI-TOF MS yeast identification in the routine clinical laboratory. Med Mycol 51:371-7.
9. Dhiman N, Hall L, Wohlfiel SL, Buckwalter SP, Wengenack NL. 2011. Performance and cost analysis of matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine identification of yeast. J Clin Microbiol 49:1614-6.

10. Marklein G, Josten M, Klanke U, Muller E, Horre R, Maier T, Wenzel T, Kostrzewa M, Bierbaum G, Hoerauf A, Sahl HG. 2009. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. J Clin Microbiol 47:2912-7.

11. Helm D, Labischinski H, Naumann D. 1991. Elaboration of a procedure for identification of bacteria using Fourier-transform IR spectral libraries: a stepwise correlation approach. Journal of microbiological methods 14:127-142.

12. Maquelin K, Kirschner C, Choo-Smith L-P, Ngo-Thi N, Van Vreeswijk T, Stämmler M, Endtz H, Bruining H, Naumann D, Puppels G. 2003. Prospective study of the performance of vibrational spectroscopies for rapid identification of bacterial and fungal pathogens recovered from blood cultures. Journal of clinical microbiology 41:324-329.

13. Brennan PJ, Nikaido H. 1995. The envelope of mycobacteria. Annual review of biochemistry 64:29-63.

14. Daffé M, Draper P. 1997. The envelope layers of mycobacteria with reference to their pathogenicity, p 131-203, Advances in microbial physiology, vol 39. Elsevier.

15. Daffé M. 2005. The cell envelope of corynebacteria. Handbook of Corynebacterium glutamicum:121-148.

16. Devriese L, Pot B. 1995. The genus enterococcus, p 327-367, The genera of lactic acid bacteria. Springer.

17. Rebuffo-Scheer CA, Schmitt J, Scherer S. 2007. Differentiation of Listeria monocytogenes serovars by using artificial neural network analysis of Fourier-transformed infrared spectra. Applied and environmental microbiology 73:1036-1040.

18. Davis R, Mauer LJ. 2011. Subtyping of Listeria monocytogenes at the haplotype level by Fourier transform infrared (FT-IR) spectroscopy and multivariate statistical analysis. International journal of food microbiology 150:140-149.

19. Romanolo K, Gorski L, Wang S, Lauzon C. 2015. Rapid identification and classification of Listeria spp. and serotype assignment of Listeria monocytogenes using Fourier transform-infrared spectroscopy and artificial neural network analysis. PloS one 10:e0143425.

20. Burckhardt I, Sebastian K, Mauder N, Kostrzewa M, Burckhardt F, Zimmermann S. 2019. Analysis of Streptococcus pneumoniae using Fourier-transformed infrared spectroscopy allows prediction of capsular serotype. European Journal of Clinical Microbiology & Infectious Diseases 38:1883-1890.

21. Marín M, Cercenado E, Sánchez-Carrillo C, Ruiz A, Gómez González Á, Rodríguez-Sánchez B, Bouza E. 2017. Accurate differentiation of Streptococcus pneumoniae from other species within the Streptococcus mitis group by peak analysis using MALDI-TOF MS. Frontiers in Microbiology 8:698.

22. Guliev RR, Suntsova AY, Vostrikova TY, Shchegolikhin AN, Popov DA, Guseva MA, Shevelev AB, Kurochkin IN. 2020. Discrimination of Staphylococcus aureus Strains from Coagulase-Negative Staphylococci and Other Pathogens by Fourier Transform Infrared Spectroscopy. Analytical Chemistry 92:4943-4948.

23. Loonen AJ, Jansz AR, Bergland JN, Valkenburg M, Wolffs PF, van den Brule AJ. 2012. Comparative study using phenotypic, genotypic, and proteomics methods for identification of coagulase-negative staphylococci. Journal of clinical microbiology 50:1437-1439.

24. Chatzigeorgiou K-S, Siafakas N, Petinaki E, Argyropoulou A, Tarpatzi A, Bobola M, Paniara O, Velegraki A, Zerva L. 2010. Identification of staphylococci by Phoenix: validation of a new protocol and comparison with Vitek 2. Diagnostic microbiology and infectious disease 68:375-381.

25. Spanu T, Sanguinetti M, Ciccaglione D, D'Inzeo T, Romano L, Leone F, Fadda G. 2003. Use of the VITEK 2 system for rapid identification of clinical isolates of staphylococci from bloodstream infections. Journal of Clinical Microbiology 41:4259-4263.

26. Dupont C, Sivadon-Tardy V, Bille E, Dauphin B, Beretti J, Alvarez A, Degand N, Ferroni A, Rottman M, Herrmann J. 2010. Identification of clinical coagulase-negative staphylococci, isolated in microbiology laboratories, by matrix-assisted laser desorption/ionization-time of flight mass spectrometry and two automated systems. Clinical Microbiology and Infection 16:998-1004.

27. Grant CE, Sewell DL, Pfaller M, Bumgardner RV, Williams JA. 1994. Evaluation of two commercial systems for identification of coagulase-negative staphylococci to species level. Diagnostic microbiology and infectious disease 18:1-5.

28. Becker K, Heilmann C, Peters G. 2014. Coagulase-negative staphylococci. Clinical microbiology reviews 27:870-926.

29. Zadoks RN, Watts JL. 2009. Species identification of coagulase-negative staphylococci: genotyping is superior to phenotyping. Veterinary microbiology 134:20-28.

30. Gee JE, De BK, Levett PN, Whitney AM, Novak RT, Popovic T. 2004. Use of 16S rRNA gene sequencing for rapid confirmatory identification of Brucella isolates. Journal of clinical microbiology 42:3649-3654.

31. Ibrahim A, Gerner-Smidt P, Liesack W. 1997. Phylogenetic relationship of the twenty-one DNA groups of the genus Acinetobacter as revealed by 16S ribosomal DNA sequence analysis. International Journal of Systematic and Evolutionary Microbiology 47:837-841.

32. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology 45:2761-2764.

33. Poyart C, Quesne G, Boumaila C, Trieu-Cuot P. 2001. Rapid and accurate species-level identification of coagulase-negative staphylococci by using the sodA gene as a target. Journal of clinical microbiology 39:4296-4301.

34. Hwang SM, Kim MS, Park KU, Song J, Kim E-C. 2011. Tuf gene sequence analysis has greater discriminatory power than 16S rRNA sequence analysis in

identification of clinical isolates of coagulase-negative staphylococci. Journal of clinical microbiology 49:4142-4149.

35. Abdul-Aziz A, Mohamad SAS, Abdullah MFF. 2015. Identification of Coagulase-Negative Staphylococci by SodA Gene Sequence Analysis. Current Research in Bacteriology 8:48.

36. Poyart C, Quesne G, Coulon S, Berche P, Trieu-Cuot P. 1998. Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase. Journal of clinical microbiology 36:41-47.

37. Poyart C, Quesnes G, Trieu-Cuot P. 2000. Sequencing the gene encoding manganese-dependent superoxide dismutase for rapid species identification of enterococci. Journal of Clinical Microbiology 38:415-418.

38. Govan JR, Deretic V. 1996. Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia. Microbiological reviews 60:539-574.

39. Pritt B, O'Brien L, Winn W. 2007. Mucoid Pseudomonas in cystic fibrosis. American journal of clinical pathology 128:32-34.

40. Marty N, Dournes J-L, Chabanon G, Montrozier H. 1992. Influence of nutrient media on the chemical composition of the exopolysaccharide from mucoid and non-mucoid Pseudomonas aeruginosa. FEMS microbiology letters 98:35-44.

41. Friedman L, Kolter R. 2004. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. Molecular microbiology 51:675-690.

42. Jackson KD, Starkey M, Kremer S, Parsek MR, Wozniak DJ. 2004. Identification of psl, a locus encoding a potential exopolysaccharide that is essential for Pseudomonas aeruginosa PAO1 biofilm formation. Journal of bacteriology 186:4466-4475.

43. Matsukawa M, Greenberg E. 2004. Putative exopolysaccharide synthesis genes influence Pseudomonas aeruginosa biofilm development. Journal of bacteriology 186:4449-4456.

44. Toyofuku M, Roschitzki B, Riedel K, Eberl L. 2012. Identification of proteins associated with the Pseudomonas aeruginosa biofilm extracellular matrix. Journal of proteome research 11:4906-4915.

45. Deshmukh DG, Zade AM, Ingole KV, Mathai JK. 2013. State of the globe: non-fermenting gram-negative bacilli challenges and potential solutions. Journal of global infectious diseases 5:125.

46. Winn W, Allen S, Janda W, Koneman E, Procop G, Schreckenberger P, Woods G. 2006. The non-fermentative Gram-negative bacilli. Color atlas and textbook of diagnostic microbiology 6.

47. Cloud JL, Harmsen D, Iwen PC, Dunn JJ, Hall G, LaSala PR, Hoggan K, Wilson D, Woods GL, Mellmann A. 2010. Comparison of traditional phenotypic identification methods with partial 5′ 16S rRNA gene sequencing for species-level identification of nonfermenting Gram-negative bacilli. Journal of clinical microbiology 48:1442-1444.

48. Baldauf NA, Rodriguez-Romo LA, Männig A, Yousef AE, Rodriguez-Saona LE. 2007. Effect of selective growth media on the differentiation of Salmonella enterica serovars by Fourier-transform mid-infrared spectroscopy. Journal of microbiological methods 68:106-114.

49. Kim S, Reuhs B, Mauer L. 2005. Use of Fourier transform infrared spectra of crude bacterial lipopolysaccharides and chemometrics for differentiation of Salmonella enterica serotypes. Journal of applied microbiology 99:411-417.

50. Kümmerle M, Scherer S, Seiler H. 1998. Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy. Appl Environ Microbiol 64:2207-2214.

51. Maquelin K, Kirschner C, Choo-Smith L-P, van den Braak N, Endtz HP, Naumann D, Puppels G. 2002. Identification of medically relevant microorganisms by vibrational spectroscopy. Journal of microbiological methods 51:255-271.

52. Bazaka K, Crawford RJ, Nazarenko EL, Ivanova EP. 2011. Bacterial extracellular polysaccharides, p 213-226, Bacterial adhesion. Springer.

53. Martin C, LiLow W, Gupta A, Cairul Iqbal Mohd Amin M, Radecka I, T Britland S, Raj P. 2015. Strategies for antimicrobial drug delivery to biofilm. Current pharmaceutical design 21:43-66.

54. Cescutti P, Bosco M, Picotti F, Impallomeni G, Leitao JH, Richau JA, Sá-Correia I. 2000. Structural study of the exopolysaccharide produced by a clinical isolate of Burkholderia cepacia. Biochemical and biophysical research communications 273:1088-1094.

55. Sawana A, Adeolu M, Gupta RS. 2014. Molecular signatures and phylogenomic analysis of the genus Burkholderia: proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus Paraburkholderia gen. nov. harboring environmental species. Frontiers in genetics 5:429.

56. Suárez-Moreno ZR, Caballero-Mellado J, Coutinho BG, Mendonça-Previato L, James EK, Venturi V. 2012. Common features of environmental and potentially beneficial plant-associated Burkholderia. Microbial ecology 63:249-266.

57. Gee JE, Glass MB, Lackner G, Helsel LO, Daneshvar M, Hollis DG, Jordan J, Morey R, Steigerwalt A, Hertweck C. 2011. Characterization of Burkholderia rhizoxinica and B. endofungorum isolated from clinical specimens. PLoS One 6:e15731.

58. Coutinho CP, Sá-Correia I, Lopes JA. 2009. Use of Fourier transform infrared spectroscopy and chemometrics to discriminate clinical isolates of bacteria of the Burkholderia cepacia complex from different species and ribopatterns. Analytical and bioanalytical chemistry 394:2161-2171.

59. Abbott SL. 2011. Klebsiella, enterobacter, citrobacter, serratia, plesiomonas, and Other Enterobacteriaceae, p 639-657, Manual of Clinical Microbiology, 10th Edition. American Society of Microbiology.

60. Bosshard P, Zbinden R, Abels S, Böddinghaus B, Altwegg M, Böttger E. 2006. 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-

GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. Journal of clinical microbiology 44:1359-1366.

61. Mignard S, Flandrois J-P. 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. Journal of microbiological methods 67:574-581.

62. Marklein G, Josten M, Klanke U, Müller E, Horre R, Maier T, Wenzel T, Kostrzewa M, Bierbaum G, Hoerauf A. 2009. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. Journal of clinical microbiology 47:2912-2917.

63. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SyN, Wang Q, Reeves PR, Wang L. 2008. Structure and genetics of Shigella O antigens. FEMS microbiology reviews 32:627-653.

64. Ewing W. 1986. The genus Escherichia. Identification of Enterobacteriaceae.

65. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. 2017. Identification of Escherichia coli and Shigella species from whole-genome sequences. Journal of clinical microbiology 55:616-623.

66. Anderson M, Sansonetti PJ, Marteyn BS. 2016. Shigella diversity and changing landscape: insights for the twenty-first century. Frontiers in cellular and infection microbiology 6:45.

67. UniProt Consortium T. 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Research 46:2699-2699.

68. Brahmbhatt H, Lindberg A, Timmis K. 1992. Shigella lipopolysaccharide: structure, genetics, and vaccine development. Pathogenesis of shigellosis:45-64.

69. Ragupathi ND, Sethuvel DM, Inbanathan F, Veeraraghavan B. 2018. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. New microbes and new infections 21:58-62.

70. (ed). 2018. Salmonella. CRC Press, Taylor & Francis Group, Roca Raton, FL. https://nls.ldls.org.uk/welcome.html?ark:/81055/vdc_100060323546.0x000001. Accessed

71. Reeves PR, Cunneen MM, Liu B, Wang L. 2013. Genetics and evolution of the Salmonella galactose-initiated set of O antigens. PLoS One 8:e69306.

72. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SyN, Reeves PR, Wang L. 2014. Structural diversity in Salmonella O antigens and its genetic basis. FEMS microbiology reviews 38:56-89.

73. Octavia S, Lan R. 2014. The Family Enterobacteriaceae, p 225-286. *In* Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (ed), The Prokaryotes: Gammaproteobacteria doi:10.1007/978-3-642-38922-1_167. Springer Berlin Heidelberg, Berlin, Heidelberg.

74. Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proceedings of the National Academy of Sciences 97:10567-10572.

75. Løbersli I, Wester AL, Kristiansen Å, Brandal LT. 2016. Molecular differentiation of Shigella spp. from enteroinvasive E. coli. European Journal of Microbiology and Immunology.

76. Brenner DJ, Fanning G, Steigerwalt A, Ørskov I, Ørskov F. 1972. Polynucleotide sequence relatedness among three groups of pathogenic Escherichia coli strains. Infection and immunity 6:308-315.

77. Martiny D, Busson L, Wybo I, El Haj RA, Dediste A, Vandenberg O. 2012. Comparison of the Microflex LT and Vitek MS systems for routine identification of bacteria by matrix-assisted laser desorption ionization–time of flight mass spectrometry. Journal of clinical microbiology 50:1313-1325.

78. Chen L, Cai Y, Zhou G, Shi X, Su J, Chen G, Lin K. 2014. Rapid Sanger sequencing of the 16S rRNA gene for identification of some common pathogens. PloS one 9:e88886.

Connecting statement 3

The developed ATR-FTIR spectral database was constructed based on spectral similarity of various groups of spectra and multiple support vector machine models for bacterial identification based on spectral fingerprints. The following chapter examines the same spectrotyping spectral database construction technique for the development of a yeast specific ATR-FTIR spectral database.

CHAPTER 5. DEVELOPMENT OF SPECTROTYPING METHOD FOR CLINICALLY RELEVANT YEASTS

5.1. ABSTRACT

Invasive fungal infections by opportunistic yeasts have increased concomitantly with the growth of an immunocompromised patient population. Misidentification of yeasts can lead to inappropriate antifungal treatment and complications. Attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy is a promising method for rapid and accurate identification of microorganisms. ATR-FTIR spectroscopy is a standalone, inexpensive, reagent-free technique that provides results within minutes after initial culture. In this study, a comprehensive spectral reference database of 65 clinically relevant yeast species was constructed and tested prospectively on spectra recorded (from colonies taken from culture plates) for 318 routine yeasts isolated from various body fluids and specimens received from 38 microbiology laboratories over a 4-month period in our clinical laboratory. ATR-FTIR spectroscopy attained comparable identification performance with matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS). In a preliminary validation of the ATR-FTIR method, correct identification rates of 100% and 95.6% at the genus and species levels, respectively, were achieved, with 3.5% unidentified and 0.9% misidentified. By expanding the number of spectra in the spectral reference database for species for which isolates could not be identified or had been misidentified, we were able to improve identification at the species level to 99.7%. Thus, ATR-FTIR spectroscopy provides a new standalone method that can rival MALDI-TOF MS for the accurate identification of a broad range of medically important yeasts. The simplicity of the ATR-FTIR spectroscopy workflow favors its use in clinical laboratories for timely and low-cost identification of life-threatening yeast strains for appropriate treatment.

5.2. INTRODUCTION

Invasive fungal infections (IFIs) by opportunistic yeasts in humans have increased over the years largely due to the concomitant growth of an immunocompromised patient population (1). Although *Candida albicans* is the leading cause of IFIs, emerging rare and non-*albicans Candida* (NAC) species are on the rise (1, 2). *Candida albicans* accounts

for approximately 50% of all IFIs followed by NAC: *Candida glabrata*, *Candida parapsilosis*, *Candida tropicalis* and *Candida krusei*, in decreasing order of frequent infection (1, 3, 4). Other yeasts like *Cryptococcus* spp., *Rhodotorula* spp., *Trichosporon* spp. and *Saccharomyces* spp., are also increasingly reported to cause infections, but are far less common (5). The problem of IFIs is compounded furthermore, with the spread of multidrug resistant yeasts, such as *Candida auris*, which has emerged rapidly worldwide and now poses a threat to public health (6-8). The correct identification of those rare and emerging yeast species is necessary for adequate antifungal therapy. The latter can be challenging with conventional identification systems used in most clinical microbiology laboratories.

Routine identification of clinical yeasts most often relies on the use of manual or automated commercial carbon assimilation identification systems such as the API 20C gallery (bioMérieux, Marcy-l'Étoile), VITEK® 2 ID/AST (bioMérieux, Marcy-l'Étoile) and Phoenix 100 ID/AST (Becton Dickinson, MD), which are used in conjunction with conventional biochemical and phenotypic assays. While gene sequencing methods are considered the gold standard and the future of infection control in clinical microbiology, they are costly, time consuming and not readily available for routine identification in most clinical sites (9). More recently, matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) has proven to be the most rapid and cost-efficient method for the identification of rare and cryptic yeast species (5, 10, 11).

Fourier transform infrared (FTIR) spectroscopy is a well-established analytical technology that has been used in various fields of study, including chemistry, forensics, ecology, medicine, astronomy and more (12-14). In microbiology, the infrared (IR) spectrum for microorganisms is representative of all compounds comprising the microorganism, and as such, IR spectroscopy is referred to as a whole-organism fingerprinting technique (15).

The acquisition of infrared spectra of microorganisms can be simplified using an attenuated total reflectance (ATR) accessory to acquire IR spectra. The IR beam from the IR source is launched at a defined angle (known as the critical angle) into an IR transparent crystal (e.g., ZnSe, Ge, Si or a diamond) whereby, total internal reflection

occurs within the crystal forming an evanescent wave (perpendicular to the propagating IR beam) above the sampling surface (Figure 5.1). Partial attenuation of the evanescent wave produced by the sample placed on the crystal takes place (Figure 5.1). An IR absorption spectrum is achieved by taking the ratio between the spectrum of the sample and the spectrum of the crystal in the absence of the sample (14).

A number of studies have demonstrated the capabilities of FTIR spectroscopy for clinical bacterial identification with promising results up to the species and strain-level, including having potential for antimicrobial susceptibility determination (16-20). For fungal identification, utilizing FTIR spectroscopy has mostly been restricted to applications in food and environmental microbiology rather than clinical microbiology (21, 22). A few studies have investigated the use of FTIR spectroscopy for the discrimination of yeast species, but those have relied on limited number of clinical isolates and species and were mostly aimed towards a general exploration of FTIR technology for identification purposes (13, 15, 23, 24).

The objective of this prospective study was to evaluate the sensitivity and specificity of an ATR-FTIR spectroscopic-based method for routine identification of medically important clinical yeasts. This necessitates construction of a comprehensive ATR-FTIR reference spectral database of clinically relevant yeasts. Identification of routine yeast samples was based on the spectral similarity between the clinical isolates to the reference spectra in the ATR-FTIR spectral database. The ATR-FTIR based method was validated prospectively by acquiring ATR-FTIR spectra of clinical yeasts obtained from 38 hospitals. The overall performance of our ATR-FTIR spectroscopic method was then compared to results obtained by the MALDI-TOF MS identification system of all clinical isolates.

5.3. MATERIALS AND METHODS

5.3.1. Construction of the ATR-FTIR reference spectral database of clinically relevant yeast species

One hundred and ninety-nine pure yeast isolates stored in 10% glycerol at -80°C were obtained from the Laboratoire de Santé Publique du Québec (LSPQ). The identification of all isolates was confirmed by gene sequencing and/or MALDI-TOF MS at

LSPQ by employing rDNA D1/D2 or ITS sequencing (using NL1-NL4 ITS1-ITS4 primers respectively) regions by comparing sequence similarity to that of sequences in Genbank, ISHAM ITS and Westerdjik Fungal Biodiversity Institute's nucleotide databases. Each sample was thawed and subcultured onto Sabouraud dextrose agar (SAB) (BD Difco, Franklin Lakes, NJ) and incubated at 30°C for 48 h followed by subculturing using the same parameters prior to spectral acquisition. Samples that displayed varying morphologies on culture plates were omitted from the study due to suspected contamination or being a mixed culture to avoid identification errors. An initial spectral database of commonly obtained routine opportunistic yeast species was constructed using a minimum of 8 or more isolates (e.g., *C. albicans*, *C. glabrata*, *C. parapsilosis, C. tropicalis*). For uncommon species (e.g., *C. auris*, *Candida haemulonii*, *Cryptococcus laurentii*, *Trichosporon inkin)* a minimum of 1-6 isolates per species were added to the initial spectral database. Subsequently, an expanded spectral reference database was created comprising a total of 789 ATR-FTIR spectra (263 isolates) from 65 species belonging to 12 genera of yeasts (Table A.9).

5.3.2. Identification of routine clinical yeast isolates

Three hundred and eighteen fresh routine clinical isolates (from skin, blood, urine and others) were collected from 38 clinical microbiology laboratories over a 4-month period and sent to LSPQ for identification by MALDI-TOF MS (VITEK MS, bioMérieux, Marcy-l'Étoile, France) using the clinical knowledge database (V3.0). These samples are independent from the created spectral database and were simultaneously identified by the ATR-FTIR based method developed in this study. All routine isolates identified by MALDI-TOF MS and ATR-FTIR spectroscopy were acquired from colonies obtained from the same agar plate. When a sample is suspected of contamination, it was subcultured and the ATR-FTIR spectra were reacquired in triplicate and reanalyzed by MALDI-TOF MS. In the case of discordant results between the ATR-FTIR and MALDI-TOF MS identification, the samples were further analyzed by rDNA D1/D2 or ITS sequencing using the same primers as mentioned above.

5.3.3. ATR-FTIR spectral acquisition

A single colony was isolated using a sterile disposable loop and deposited directly onto the sampling surface of an ATR-FTIR spectrometer (Cary 630, Agilent Technologies, Santa Clara, CA) (Figure 5.1). The spectra were acquired using a spectral resolution of 8 cm^{-1} with 64 co-additions for the background and sample scan in the spectral range between 4000 and 650 cm^{-1}. After spectral acquisition, disinfection of the sampling surface was achieved by wiping the ATR surface with lint-free paper moistened with 70% ethanol. Triplicate spectra were acquired from three individual colonies per agar plate to demonstrate spectral reproducibility and sample purity. The triplicate spectra were employed for the construction of the reference spectral database or the identification of clinical routine yeast isolates.

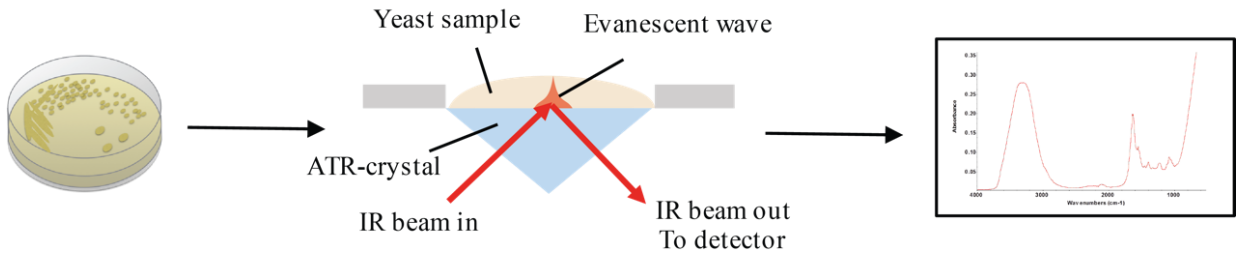


Figure 5.1. Experimental workflow of the ATR-FTIR spectroscopic-based method for yeast analysis. A single colony is directly transferred from the agar plate (without prior treatment) onto the ATR sampling surface of the ATR-FTIR spectrometer. An infrared (IR) beam is directed into the ATR crystal resulting in the generation of an evanescent wave perpendicular to the propagating IR beam within the crystal. Attenuation of the evanescent wave by the sample yields the ATR-FTIR spectrum in ~1 minute. Identification of the sample is based on its spectral similarity to that of an infrared spectrum of an isolate in the ATR-FTIR reference spectral database.

5.3.4. Spectral quality assessment, processing and analysis

All ATR-FTIR spectra underwent a spectral quality check (prior to incorporation into the spectral reference database or for identification) by employing hierarchical cluster analysis (HCA) as a measure of similarity between the triplicate spectra (i.e., how close they cluster together in a dendrogram comprising spectra from the same species). The spectra were then subjected to vector normalization and their 1st derivative computed by

an in-house written software or with commercially available spectral analysis software (OMNIC, Thermo Fisher Scientific, Madison, WI). Principal component analysis (PCA) was used in conjunction with a forward region selection algorithm to discriminate among classes (genera, species) based on specific spectral regions in the ATR-FTIR spectra for each class (25).

5.3.5. ATR-FTIR spectroscopy-based identification strategy

The reference spectral database was constructed with 789 spectra acquired from 65 species belonging to 12 genera. Hierarchical database structure was employed, whereby identification is performed in two stages: (i) classification at the genus level, and (ii) at the species level, employing spectral features appropriate for differentiating between strains at each stage of the process. Using the same approach, the identification of the spectral regions for the discrimination between the different genera and species (within each genus) is carried out in two stages: a grid search followed by a “greedy” search, with the grid search filtering through large spectral regions and providing starting points for the greedy. The details of this algorithm are beyond the scope of this article and is presented elsewhere (26).

Differentiation between the species within each genus is based on spectral differences and similarities (reflecting differences in the biochemical and metabolic profiles) of the microorganisms. A multitier spectral database was constructed using pair-wise spectral groups representative of different genera or species within a genus. The pair-wise grouping in each tier is based on the relative spectral distance computed by PCA. In the identification step, the ATR-FTIR spectrum of a routine isolate is assigned to either group within a given tier or, flagged as an outlier based on spectral dissimilarity to the spectra in the pair-wise group and is not identified. This process is iterated in a stepwise fashion for all tiers until identification (or no identification) of the isolate at the species-level is achieved. Comparable results can also be achieved with the use of the discriminate analysis routines in commercially available multivariate statistical software, JMP® Pro (SAS, Cary, NC). Details of the ATR-FTIR spectral database structure can be found in Figure 5.2.

5.4. RESULTS

5.4.1. Identification of clinical yeasts by ATR-FTIR spectroscopy

Using a preliminary ATR-FTIR spectral reference database of 199 isolates belonging to 5 genera and 14 species, the database yielded a 95.6% correct species identification (*n*=304) of 318 routine clinical yeast isolates (in concordance with MALDI-TOF MS and rDNA D1/D2 sequencing). One hundred percent correct species identification was achieved for all routine samples belonging to *Candida dubliniensis*, *C. krusei*, *Candida lusitaniae*, *Candida orthopsilosis*, *C. parapsilosis*, *Cryptococcus neoformans*, *Rhodotorula mucilaginosa* and *Saccharomyces cerevisiae*. The remaining 4.4% (*n*=14) routine isolates were either unidentified (*n*=11) due to the lack of representation in the reference spectral database or misidentified (*n*=3) (Table 5.1).

5.4.2. Expansion and re-validation of the ATR-FTIR spectral reference database

To address the limited representation of clinical yeasts from additional genera and species, the ATR-FTIR spectral reference database was expanded by adding spectra from 7 new genera and 51 species (Table A.9). Re-analysis of the spectra from the 318 isolates with the expanded spectral database yielded a 99.7% correct species identification of routine clinical isolates (Table 5.1). It should be noted that one of the isolates identified by ATR-FTIR spectroscopy as *Meyerozyma caribbica,* was discordant with MALDI-TOF MS which identified it as *Candida guilliermondii* with 99.9% confidence. This sample was re-cultured a second time and yielded the same results by both ATR-FTIR spectroscopy and MALDI-TOF MS analysis. The isolate was identified by rDNA D1/D2 sequencing (NL1-NL4 primers) as *M. caribbica.* Ultimately, one isolate, *C. tropicalis*, was misidentified as *C. lusitaniae* (confirmed by rDNA D1/D2 sequencing) resulting in an overall misidentification rate of 0.3%.

The spectra acquired from the 318 isolates were further added to the expanded database and employed in an external prospective validation of an additional 143 isolates to re-validate the newly enhanced spectral database (comprising a total of 1743 spectra) yielding a 100% correct species level of identification of 142 isolates represented in the spectral database (Table 5.2). One isolate remained unidentified yielding an overall

99.3% correct identification. The identity of the unidentified isolate was established as *Saprochaete clavata* by gene sequencing. It is of interest to note that MALDI-TOF MS also provided no identification of the unidentified isolate due its absence in the VITEK MS database (V3.0).

Table 5.1. Identification of routine clinical yeast isolates obtained using a preliminary ATR-FTIR spectral reference database

	No. (%) of isolates			
	Routine isolates			
Microorganism	Total	Identified	Unidentified	Misidentified
Candida spp.				
C. albicans	130	128 (98.5)	0	2
C. dubliniensis	13	13	0	0
C. glabrata	68	67 (98.5)	1	0
C. guilliermondii	2	0 (0.00)	2[a]	0
C. kefyr	2	1 (50.0)	1	0
C. krusei	8	8	0	0
C. lipolytica	1	0 (0.00)	1	0
C. lusitaniae	16	16	0	0
C. orthopsilosis	4	4	0	0
C. parapsilosis	32	32	0	0
C. pararugosa	1	0 (0.00	1	0
C. pelliculosa	1	0 (0.00)	1	0
C. tropicalis	23	21 (91.3)	1	1
C. utilis	3	0 (0.00)	3	0
Cryptococcus neoformans	7	7	0	0
Rhodotorula mucilaginosa	1	1	0	0
Saccharomyces cerevisiae	5	5	0	0
Trichosporon spp.	1	1 [b]	0	0
Total	318	304 (95.6)	11 (3.5)	3 (0.9)

[a]1 of the 2 unidentified isolate of *C. guilliermondii* was later identified as *Meyerozyma caribbica* by gene sequencing. This isolate was omitted from calculation.

[b]*Trichosporon* spp. are underrepresented with less than 5 isolates per species. The isolate correctly identified at the genus level and later identified as *Trichosporon mycotoxinivorans* by rDNA sequencing.

Table 5.2. Identification results obtained with the expanded spectral database

	No. (%) of isolates			
	Routine isolates			
Microorganism	Total	Identified	Unidentified	Misidentified
Candida spp.				
C. albicans	130	130	0	0
C. dubliniensis	13	13	0	0
C. glabrata	68	68	0	0
C. guilliermondii	1	1	0	0
C. kefyr	2	2	0	0
C. krusei	8	8	0	0
C. lipolytica	1	1	0	0
C. lusitaniae	16	16	0	0
C. orthopsilosis	4	4	0	0
C. parapsilosis	32	32	0	0
C. pararugosa	1	1	0	0
C. pelliculosa	1	1	0	0
C. tropicalis	23	22 (95.7)	0	1[a]
C. utilis	3	3	0	0
Cryptococcus neoformans	7	7	0	0
Meyerozyma caribbica	1	1	0	0
Rhodotorula mucilaginosa	1	1	0	0
Saccharomyces cerevisiae	5	5	0	0
Trichosporon spp.	1	1	0	0
Total	318	317 (99.7)	0	1 (0.3)

[a]Same isolate of *C. tropicalis* in Table 5.1 was misidentified as *C. lusitaniae*.

5.5. DISCUSSION

A preliminary ATR-FTIR spectral reference database of 199 isolates belonging to 5 genera and 14 species was constructed. The database (constructed with DataAnalysis and MLC) yielded a 95.6% correct species identification (*n*=304) of 318 routine clinical yeast isolates (in concordance with MALDI-TOF MS and rDNA D1/D2 sequencing). 100% correct species identification was achieved for all routine samples belonging to *C. dubliniensis, C. krusei, C, lusitaniae, C. orthopsilosis, C. parapsilosis, C. neoformans, R. mucilaginosa* and *S. cerevisiae*. The remaining 4.4% (*n*=14) routine isolates were either unidentified (*n*=11) due to the lack of representation in the reference spectral database or misidentified (*n*=3) (data not shown).

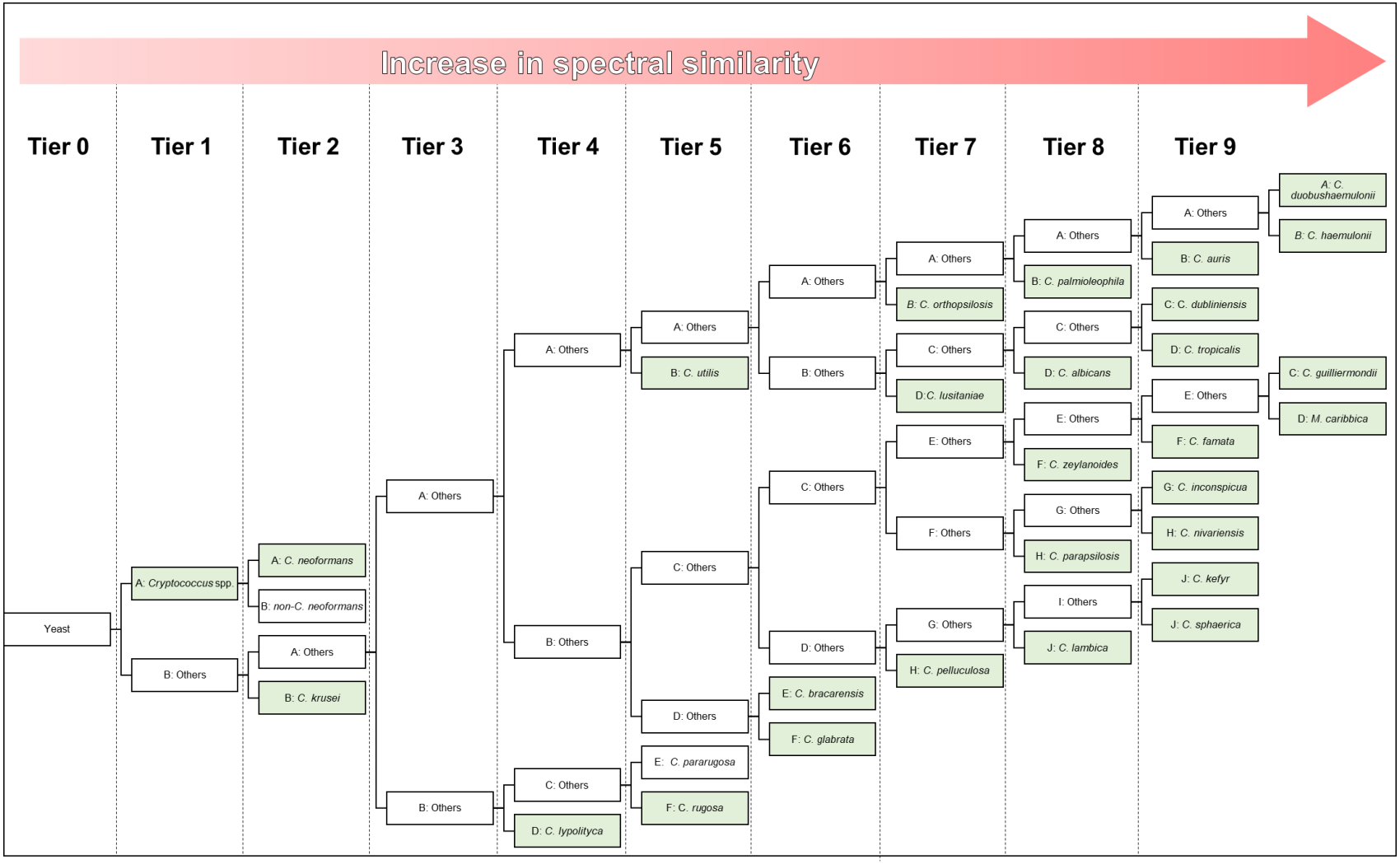


Figure 5.2. Pairwise multitier ATR-FTIR spectral database structure for the species identification of clinically relevant yeasts

To address the limited representation of clinical yeasts from additional genera and species, the ATR-FTIR spectral reference database was expanded by adding spectra from 7 new genera and 51 species. Using the expanded spectral reference database comprising of 65 species from 12 different genera, the expanded database yielded a 99.7% correct species identification of the routine clinical isolates (*n*= 318) (Table 5.2). It should be noted that one of the isolates identified by ATR-FTIR spectroscopy as *Meyerozyma caribbica* was discordant with MALDI-TOF MS which identified it as *C. guilliermondii* with 99.9% confidence. This sample was re-cultured a second time and yielded the same results by both ATR-FTIR spectroscopy and MALDI-TOF MS analysis. The isolate was identified by rDNA D1/D2 sequencing (NL1-NL4 primers) as *Meyerozyma caribbica*. Ultimately, only one isolate, *C. tropicalis*, was misidentified as *C. lusitaniae* (confirmed by rDNA D1/D2 sequencing) resulting in an overall misidentification rate of 0.3%.

To our knowledge, no prior study has utilized ATR-FTIR spectroscopy for routine identification of clinical yeasts. Our group is the first to create a comprehensive clinical yeast ATR-FTIR spectral reference database comprising of 263 reference strains of 65 species belonging to 12 genera encompassing both rare and emerging strains such as *C. auris* and *Trichosporon asahii*. As with any method used for fungal identification for diagnostic purposes, misidentification is a major concern for appropriate species-specific antifungal therapy; it is more desirable to obtain an unidentified isolate result opposed to a false identification. The higher rate of unidentified (3.5%) than misidentified (0.9%) species was observed in our initial spectral database study where it did not include enough isolates for rare species such as *Candida utilis*, *Candida pararugosa* and *Trichosporon mycotoxinivorans*. This limitation was addressed by expanding the spectral database to encompass a larger number of species from additional genera. By employing the expanded ATR-FTIR database for the analysis of 318 routine clinical isolates, 100% and 99.7% correct identification at the genus level and species was achieved respectively. Accordingly, the single-step protocol employed for the analysis of clinical isolates by ATR-FTIR spectroscopy makes it both superior and more cost effective than any currently available identification platform. For example, correct identification of medically important yeasts reported for VITEK 2, API ID32C, Phoenix ID and AuxaColor

and VITEK MS ranges from 72.7-97.1% with misidentification ranging from 0.4-33.3% (27-31). MALDI-TOF MS has overcame difficulties of identifying closely related complexed yeasts such as *C. albicans*/*C. dubliniensis*, and *C. parapsilosis*/*C. orthopsilosis* in comparison to conventional biochemical techniques (29). Likewise, ATR-FTIR-based results from this study also demonstrated that the IR spectra contained enough information to allow discrimination between *C. albicans*, *C. dubliniensis*, *C. parapsilosis* and *C. orthopsilosis*.

To improve the ATR-FTIR-based method for identification of routine isolates, the inclusion of the five species not initially present in the IR spectral database resulted in correct identification of all 7 initially unidentified isolates. Similarly, species that had initially resulted in either a misidentification or no identification, resulted in correct identification of 5 out of 6 routine isolates using the expanded spectral database. Overall, only 1 *C. tropicalis* remained erroneously identified as *C. lusitaniae* by ATR-FTIR spectroscopy with the expanded infrared spectral database. The misidentified *C. tropicalis'* colonies were typical and by visual identification is in accordance with what is known, and misidentification of the species as *C. lusitaniae* is not common and are phenotypically dissimilar – further investigation is needed. It should be noted that one isolate identified as *Meyerozyma caribbica* by rDNA sequencing, was misidentified as *C. guilliermondii* (with 99.9% confidence) using the VITEK MS system. As both species are part of same species complex and are genetically similar (32), the MALDI-TOF MS misidentification error is considered minor. This error maybe attributed to the absence of a mass spectrometric spectral representation of *M. caribbica* in the VITEK MS database (clinical knowledge database V3.0). Similarly, the initial ATR-FTIR spectral database did not include IR spectra representative of *M. caribbica,* however, unlike MALDI-TOF MS, the sample was unidentified rather than misidentified. Inclusion of IR spectra of *M. caribbica* to the expanded IR spectral database of the clinical yeast isolates resulted in its correct identification by ATR-FTIR spectroscopy. Additionally, two isolates identified as *C. lusitaniae* by rDNA sequencing, were unidentified by MALDI-TOF MS where they were both correctly identified by ATR-FTIR spectroscopy. Based on these findings, ATR-FTIR spectroscopy thus offers several advantages over MALDI-TOF MS, being an inexpensive,

reagent-free, and one-step procedure – obviating the need for consumables, or the need for an extraction step with the use of harsh acids (e.g., formic acid). In addition, the method is free from interruption associated with time delays related to drying of the chemical matrix (and acid) and achieving a high vacuum prior to MS spectral acquisition.

The minimum number of reference strains required for inclusion in the IR spectral reference database may be species specific and dependent on the variability between the biochemical composition of the entire microorganism for a given genus or species. For example, the spectral differences between *Candida pelliculosa* and the other species were substantial and spectral representation from only two *C. pelliculosa* isolates in the spectral database was sufficient for complete discrimination. Other species, such as *C. albicans* and *C. dubliniensis* were more difficult to discriminate between each other and required additional ATR-FTIR spectra resulting in >13 reference strains per species for complete discrimination. Likewise, the inclusion of ATR-FTIR spectra of newly discovered and emerging species in the IR spectral database is expected to enhance the predictive performance of the ATR-FTIR spectroscopy-based method. Furthermore, inclusion of the correctly identified routine samples in this study to the expanded database should increase its specificity for future analysis of routine clinical isolates.

5.6. CONCLUSION

Overall, our central study demonstrates a strong potential for ATR-FTIR spectroscopy for rapid-routine analysis of clinical yeasts with an overall 99.7% correct species-level identification. ATR-FTIR spectroscopy demonstrated advantages over current conventional biochemical, MALDI-TOF MS and gene sequencing identification methods. ATR-FTIR spectroscopy requires no sample preparation after incubation, is a standalone method, as little as a single colony is needed, is reagent-free and complete data acquisition and analysis for the identification at the species level is completed in less than 2 minutes per sample. These advantages make our ATR-FTIR-based method the fastest and lowest costing technology developed to date for microorganism identification.

5.7. REFERENCES

1. Miceli MH, Díaz JA, Lee SA. 2011. Emerging opportunistic yeast infections. The Lancet Infectious Diseases 11:142-151.

2. Vazquez-Gonzalez D, Perusquia-Ortiz AM, Hundeiker M, Bonifaz A. 2013. Opportunistic yeast infections: candidiasis, cryptococcosis, trichosporonosis and geotrichosis. J Dtsch Dermatol Ges 11:381-93; quiz 394.

3. Maquelin K, Kirschner C, Choo-Smith LP, Ngo-Thi NA, van Vreeswijk T, Stammler M, Endtz HP, Bruining HA, Naumann D, Puppels GJ. 2003. Prospective Study of the Performance of Vibrational Spectroscopies for Rapid Identification of Bacterial and Fungal Pathogens Recovered from Blood Cultures. Journal of Clinical Microbiology 41:324-329.

4. McCarty TP, Pappas PG. 2016. Invasive Candidiasis. Infect Dis Clin North Am 30:103-24.

5. Dhiman N, Hall L, Wohlfiel SL, Buckwalter SP, Wengenack NL. 2011. Performance and cost analysis of matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine identification of yeast. J Clin Microbiol 49:1614-6.

6. Ben-Ami R, Berman J, Novikov A, Bash E, Shachor-Meyouhas Y, Zakin S, Maor Y, Tarabia J, Schechner V, Adler A, Finn T. 2017. Multidrug-Resistant Candida haemulonii and C. auris, Tel Aviv, Israel. Emerg Infect Dis 23.

7. Spivak ES, Hanson KE. 2018. Candida auris: an Emerging Fungal Pathogen. Journal of Clinical Microbiology 56:1-10.

8. Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, Colombo AL, Calvo B, Cuomo CA, Desjardins CA, Berkow EL, Castanheira M, Magobo RE, Jabeen K, Asghar RJ, Meis JF, Jackson B, Chiller T, Litvintseva AP. 2017. Simultaneous Emergence of Multidrug-Resistant Candida auris on 3 Continents Confirmed by Whole-Genome Sequencing and Epidemiological Analyses. Clin Infect Dis 64:134-140.

9. Raoult D, Fournier PE, Drancourt M. 2004. What does the future hold for clinical microbiology? Nat Rev Microbiol 2:151-9.

10. Cassagne C, Cella AL, Suchon P, Normand AC, Ranque S, Piarroux R. 2013. Evaluation of four pretreatment procedures for MALDI-TOF MS yeast identification in the routine clinical laboratory. Med Mycol 51:371-7.

11. Marklein G, Josten M, Klanke U, Muller E, Horre R, Maier T, Wenzel T, Kostrzewa M, Bierbaum G, Hoerauf A, Sahl HG. 2009. Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. J Clin Microbiol 47:2912-7.

12. Nikolić GS. 2011. Fourier Transforms - New Analytical Approaches and FTIR Strategies. InTech, Janeza Trdine 9, 51000 Rijeka, Croatia.

13. Santos C, Fraga ME, Kozakiewicz Z, Lima N. 2010. Fourier transform infrared as a powerful technique for the identification and characterization of filamentous fungi and yeasts. Res Microbiol 161:168-75.

14. Smith BC. 2001. Fundamentals of Fourier Transform Infrared Spectroscopy, Second Edition ed. Taylor and Francis Group, LLC, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742.

15. Silva S, Tobaldini-Valerio F, Costa-de-Oliveira S, Henriques M, Azeredo J, Ferreira EC, Lopes JA, Sousa C. 2016. Discrimination of clinically relevant Candida species by Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR-ATR). RSC Advances 6:92065-92072.

16. Bosch A, Minan A, Vescina C, Degrossi J, Gatti B, Montanaro P, Messina M, Franco M, Vay C, Schmitt J, Naumann D, Yantorno O. 2008. Fourier transform infrared spectroscopy for rapid identification of nonfermenting gram-negative bacteria isolated from sputum samples from cystic fibrosis patients. J Clin Microbiol 46:2535-46.

17. Sandt C, Madoulet C, Kohler A, Allouch P, De Champs C, Manfait M, Sockalingum GD. 2006. FT-IR microspectroscopy for early identification of some clinically relevant pathogens. J Appl Microbiol 101:785-97.

18. Wenning M, Breitenwieser F, Konrad R, Huber I, Busch U, Scherer S. 2014. Identification and differentiation of food-related bacteria: A comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry. J Microbiol Methods 103:44-52.

19. Amiali NM, Golding GR, Sedman J, Simor AE, Ismail AA. 2011. Rapid identification of community-associated methicillin-resistant Staphylococcus aureus by Fourier transform infrared spectroscopy. Diagn Microbiol Infect Dis 70:157-66.

20. Amiali NM, Mulvey MR, Sedman J, Simor AE, Ismail AA. 2007. Epidemiological typing of methicillin-resistant Staphylococcus aureus strains by Fourier transform infrared spectroscopy. J Microbiol Methods 69:146-53.

21. Lecellier A, Gaydou V, Mounier J, Hermet A, Castrec L, Barbier G, Ablain W, Manfait M, Toubas D, Sockalingum GD. 2015. Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds. Food Microbiol 45:126-34.

22. Corte L, Antonielli L, Roscini L, Fatichenti F, Cardinali G. 2011. Influence of cell parameters in Fourier transform infrared spectroscopy analysis of whole yeast cells. Analyst 136:2339-49.

23. Taha M, Hassan M, Essa S, Tartor Y. 2013. Use of Fourier transform infrared spectroscopy (FTIR) spectroscopy for rapid and accurate identification of Yeasts isolated from human and animals. International Journal of Veterinary Science and Medicine 1:15-20.

24. Costa FSL, Silva PP, Morais CLM, Arantes TD, Milan EP, Theodoro RC, Lima KMG. 2016. Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii. Analytical Methods 8:7107-7115.

25. Kirkwood J, Ghetler A, Sedman J, Leclair D, Pagotto F, Austin JW, Ismail AA. 2006. Differentiation of group I and group II strains of Clostridium botulinum by focal plane array Fourier transform infrared spectroscopy. Journal of Food Protection 69:2377-2383.

26. Ghetler A. 2009. Development of an expert system for the identification of bacteria by focal plane array Fourier transform infrared spectroscopy. McGill University Library.

27. Kim TH, Kweon OJ, Kim HR, Lee MK. 2016. Identification of Uncommon Candida Species Using Commercial Identification Systems. J Microbiol Biotechnol 26:2206-2213.

28. Carvalho A, Costa-De-Oliveira S, Martins ML, Pina-Vaz C, Rodrigues AG, Ludovico P, Rodrigues F. 2007. Multiplex PCR identification of eight clinically relevant Candida species. Med Mycol 45:619-27.

29. Posteraro B, Efremov L, Leoncini E, Amore R, Posteraro P, Ricciardi W, Sanguinetti M. 2015. Are the Conventional Commercial Yeast Identification Methods Still Helpful in the Era of New Clinical Microbiology Diagnostics? A Meta-Analysis of Their Accuracy. J Clin Microbiol 53:2439-50.

30. Wang H, Fan YY, Kudinha T, Xu ZP, Xiao M, Zhang L, Fan X, Kong F, Xu YC. 2016. A Comprehensive Evaluation of the Bruker Biotyper MS and Vitek MS Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry Systems for Identification of Yeasts, Part of the National China Hospital Invasive Fungal Surveillance Net (CHIF-NET) Study, 2012 to 2013. J Clin Microbiol 54:1376-80.

31. Ling H, Yuan Z, Shen J, Wang Z, Xu Y. 2014. Accuracy of matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of clinical pathogenic fungi: a meta-analysis. J Clin Microbiol 52:2573-82.

32. Romi W, Santosh K, Giasuddin A, Kumaraswamy J. 2014. Reliable differentiation of Meyerozyma guilliermondii from Meyerozyma caribbica by internal transcribed spacer restriction fingerprinting. BMC Microbiology 14:52-62.

Connecting statement 4

The following chapter is a multicenter evaluation study between 6 clinical microbiology laboratories across Quebec. The study aims at evaluating the standardized method and ATR-FTIR spectral reference database developed in the previous chapter for the identification of clinically relevant yeasts and to discuss the limitations in hopes of advancing the technology and technique for routine use.

CHAPTER 6. MULTICENTER EVALUATION OF ATTENUATED TOTAL REFLECTANCE FOURIER TRANSFORM INFRARED SPECTROSCOPY-BASED METHOD FOR RAPID IDENTIFICATION OF CLINICALLY RELEVANT YEASTS

6.1. ABSTRACT

Fourier transform infrared (FTIR) spectroscopy has demonstrated applicability as a reagent-free whole-organism fingerprinting technique for both microbial identification and strain typing. For routine application of this technique in clinical microbiology laboratories, acquisition of FTIR spectra in the attenuated total reflectance (ATR) mode simplifies the FTIR spectroscopy workflow, providing results within minutes after initial culture without prior sample preparation. In our previous central work, 99.7% correct species identification of clinically relevant yeasts was achieved by employing an ATR-FTIR-based method and spectral database developed by our group. In this prospective study, ATR-FTIR spectrometers were placed in 6 clinical microbiology laboratories over a collective 16-month period and were used to collect spectra belonging to yeasts for identification. The identification results were compared to those obtained from conventional biochemical tests and/or a matrix-assisted laser desorption/ionization time of flight mass spectrometry obtained from the participating laboratories. Discordant results were reanalyzed by routine identification methods, ATR-FTIR spectroscopy and PCR gene sequencing of the D1D2/ITS regions. Among the 534 routine clinical yeast isolates collected and identified by the ATR-FTIR-based method, 525 (98.3%) isolates were correctly identified at the species level while the remaining isolates were inconclusive with no misidentifications. Additional randomly selected yeasts (*n*=39) and *Candida auris* (*n*=24) isolates were also evaluated and resulted in 100% correct identification. Our data suggest that ATR-FTIR spectroscopy demonstrates a reliable, cost-effective yeast identification technique that provides accurate and timely species identification promptly after the initial culture multicentrally using multiple spectrometers in different laboratories with multiple system operators.

6.2. INTRODUCTION

Fungal infections affect over a billion people worldwide, resulting in an estimated 1.5 million deaths each year (1). It is also estimated that there are approximately 700,000

global cases of invasive candidiasis and over 220,000 cases of fungal disease due to *Cryptococcus neoformans* associated with HIV/AIDs complication (2, 3). Over the past few decades, there has been a reported increase in nosocomial candidiasis where candidemia is associated with mortality rates of over 40% of those infected (4-6). Furthermore, *Candida* spp. in many developed countries are the 3rd or 4th leading cause of nosocomial bloodstream infections. These opportunistic microorganisms pose the greatest risk to the elderly population, neonates and those who are immunocompromised (6). Although *Candida albicans* accounts for over 40% of all yeast infections, there is an increasing prevalence of non-*C. albicans* species infections. For example, the increase in incidences of *Candida auris* is troublesome due to its ability to acquire antifungal resistance and is easily transmitted from person-to-person – complicating treatment and/or resulting in poor patient outcomes (6-8).

Rapid identification of yeasts is necessary for appropriate patient care and to reduce the spread of antifungal resistance. Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) has revolutionized microbial identification in both bacteriology and mycology in the past decade owing to its simplicity, rapidness, and reliability for speciating microorganisms relative to conventional biochemical techniques (9-12). It is also the first commercially available and widely accepted spectroscopic technique for *in vitro* diagnostics. Although it has revolutionized clinical microbiology, there is a high initial capital cost associated with the technique along with expenses for reagents, the requirement for additional pre-treatment steps for fungi, disposable target plates and maintenance (13), has limited the implementation of international surveillance programs aimed at tracking fungal infections.

Within the realm of molecular spectroscopy, Fourier transform infrared (FTIR) spectroscopy has been widely applied for the investigation of cell metabolism, microbial identification, and strain-typing (14-23). However, there is a lack of standardization and spectral database evaluation studies in the clinical setting. When coupled with the attenuated total reflectance (ATR) mode of spectral acquisition, ATR-FTIR spectroscopy becomes a rapid, reagent-free, and low maintenance technique that does not require

sample preparation with harsh chemicals after culture nor drying or vacuum-related downtime.

Our group, in previous work has created and developed an ATR-FTIR spectroscopy-based microbial spectrotyping (defined as the process of determining the differences between spectra obtained from microorganisms based on the absence or presence, and relative intensities of infrared absorption bands) technique for rapid identification of clinically relevant yeasts (14). Spectrotyping entails the implementation of a standardized culturing methods and well-defined ATR-FTIR spectral acquisition and preprocessing techniques to construct a spectral database of highly reproducible ATR-FTIR spectra of well characterized clinical yeasts isolates. Through spectrotyping, a database was created consisting of 65 species and 23 genera and was subsequently evaluated centrally by acquiring 318 routine yeasts isolates, resulting in 100% correct genus and 99.7% correct species identification (14).

The results of the latter study warranted the undertaking of a prospective multicenter evaluation study as a first step in the implementation of ATR-FTIR spectroscopy for routine yeast identification. The following study aims at prospectively evaluating the performance of our ATR-FTIR reference spectral database for rapid identification of routine yeasts on-site in multiple clinical microbiology laboratories. Results obtained from the ATR-FTIR spectroscopy-based identification results were compared to standard identification techniques for diagnostics (conventional biochemical tests, MALDI-TOF MS and/or rDNA gene sequencing) implemented by the participating laboratories.

6.3. MATERIALS AND METHOD

6.3.1. Sample collection

In routine, microorganisms were isolated from clinical specimens (i.e., blood, urine, respiratory specimen) onto culture agar media (i.e., Sabouraud dextrose agar (SAB), Sabouraud dextrose agar Emmons or inhibitory mold agar). A total of 534 isolates [17 genera and 29 species (Table 6.1)] were collected from Centre hospitalier universitaire de Sherbrooke (CHUS) (*n*=93), Centre hospitalier universitaire Sainte Justine (CHUSJ) (*n*=12), Laboratoire de santé publique du Québec (LSPQ) (*n*=132), Hôpital Maisonneuve-

Rosemont (HMR) (*n*=100), Centre universitaire de santé McGill (CUSM) (*n*=100) and Centre hospitalier de l'Université de Montréal (CHUM) (*n*=97) over a cumulative 16-month period.

Isolates collected from LSPQ and CHUSJ were analyzed in real-time while those collected from CHUS, HMR, CUSM and CHUM where prospectively collected and stored at -80°C and re-cultured from frozen for ATR-FTIR spectral acquisition to facilitate batch testing due to limited available spectrometers and technicians. All centers' isolates were collected to be representative of routine isolates over a given time frame to accumulate approximately 100 isolates.

CHUSJ (children's hospital) provided 12 isolates from routine over a 2-month period (in 2019). An additional 39 isolates (non-routine) were randomly collected at CHUSJ in 2020, belonging to 8 *Candida* species (Table A.10) isolated from children from sterile sites and/or that were difficult to identify in routine, were included in the evaluation study (results were not included with multicenter evaluation results). Additionally, due to the low prevalence of *C. auris* (*n*=1) in routine for the multicenter study (but high importance for accurate identification), it was necessary to obtain additional isolates of *C. auris*. Twenty-four well characterized (by whole genome sequencing and PCR) *C. auris* isolates (Clade I to IV) were provided by the Centers for Disease Control and Prevention (CDC) (Atlanta, GA).

6.3.2. Standard routine species identification and antimicrobial susceptibility testing

Isolates collected from CHUS, CUSM, CHUM, HMR and LSPQ were identified by MALDI-TOF MS (VITEK MS, bioMérieux, Marcy-l'Étoile, France) using the clinical knowledge database (V3.2), while isolates collected from CHUSJ were identified by VITEK 2 version 8.01 (bioMérieux, Marcy-l'Étoile, France) or were sent to LSPQ for definitive identification. Supplemental isolates of *C. auris* (*n*=24) received from CDC were further confirmed by MALDI-TOF MS and ITS and D1D2 rDNA gene sequencing at LSPQ.

Broth microdilution antimicrobial susceptibility testing following the Clinical and Laboratory Standards Institute guidelines M60 (Performance Standards for Antifungal

Susceptibility Testing of Yeasts) were completed at LSPQ on select isolates as part of routine and in one instance, the results of the test were used for supplemental observation for discussion.

6.3.3. Sample preparation for ATR-FTIR spectral acquisition

Following the same protocol of our previous study (14), all isolates that were not routinely cultured onto Sabouraud dextrose agar (SAB, 40 g/L dextrose, pH = 5.6±0.2 at room temperature) (BD Difco, Franklin Lakes, NJ) for real-time analysis at LSPQ and CHUS were subcultured onto SAB and incubated at 30°C for 48 h prior to ATR-FTIR spectral acquisition. All other routine isolates collected from the remaining institutions from frozen were cultured and subcultured onto SAB and incubated at 30°C for 48 h.

6.3.4. ATR-FTIR spectroscopy-based identification

Species identification by ATR-FTIR spectroscopy was achieved by employing a previously created ATR-FTIR spectral database (referred herein as the “reference database”) for identifying clinically relevant yeasts. The reference database was constructed with spectra belonging to 263 reference strains (made available by LSPQ) that encompasses 65 species belonging to 23 genera of clinically relevant yeasts (14) (Table A.9) Briefly, the reference database was constructed with averaged triplicate spectra of yeast strains. Identification of an unknown spectrum belonging to an isolate of yeast was achieved by the interrogation of a sequential multitier pairwise search reference database by comparing it’s ATR-FTIR spectrum to those in the reference spectral database (14). It should be noted that species of the genus *Trichosporon* were identified only to the genus level by the ATR-FTIR spectroscopy-based method due to low representation (less than 4) within individual species but significant representation at the genus level (n=7).

Confidences of the identification of unknowns were based on the standard deviation (SD) from the spectral similarity of the unknown spectrum from the mean spectrum of the reference spectra in the reference database (that it is being predicted as) and the closest spectral similarity match to those spectra in the reference database. A SD from the mean

spectrum of greater than 3 and low spectral similarity (<70%) indicates no identification (i.e., no spectral match between the spectrum of the unknown isolate to those represented in the reference database). Additionally, different species (or genus) identification from each of the triplicate spectra collected from one strain (grown on one SAB plate) was also considered inconclusive (i.e., no identification is reported), and possibly indicating the presence of a contaminating microorganism.

Discordant identification results between ATR-FTIR spectroscopy and standard routine identification methods were addressed by re-culturing the isolate and re-acquiring the ATR-FTIR spectra along with re-analyzing the isolate by VITEK MS (bioMérieux, Marcy-l'Étoile, France). If results remain discordant, definitive identification was achieved at the LSPQ by PCR rDNA gene sequencing of the D1/D2 (NL1 and NL4 primers (24)) and ITS regions (ITS1 and 4 primers (25)) from the same re-culture plates.

6.4. RESULTS

A total of 534 routine clinical yeast isolates were collected from 6 clinical microbiology laboratories and were identified by the ATR-FTIR-based method developed in our previous work (14). A 100% correct identification of all isolates at the genus level was achieved. A total of 525/534 (98.3%) of the isolates were correctly identified to the species level (Table 6.1). The 9 remaining isolates were not identified (i.e., results were inconclusive). Inconclusive results are those with less than 70% spectral match from the reference spectra with SD greater than 3 of the spectral similarity from the reference predicted mean spectrum of its closest spectral match. Re-culturing the isolates a second time and their analysis by MALDI-TOF MS and ATR-FTIR spectroscopy did not resolve the discrepancy. Further analysis by rDNA/gene sequencing showed agreeance with the MALDI-TOF MS results while species identification based on ATR-FTIR spectroscopy remained inconclusive. Four of the 9 isolates belonged to 4 different species (*Candida metapsilosis, Saprochaete clavata* (*Geotrichum clavatum*), *Starmerella* (*Candida*) *magnoliae*, *Torulaspora delbrueckii* (*Candida colliculosa*)) that were underrepresented in the spectral database (i.e., the database had <5 isolates present in the spectral reference database; Table 1). The 5 remaining isolates belonged to 4 species (*Candida dubliniensis* (*n*=1), *Candida glabrata* (*n*=1), *Candida orthopsilosis* (*n*=1) and *Meyerozyma*

guilliermondii (*n*=2)) which had higher spectral representation in the reference spectral database (>9 isolates per species). Thirteen isolates (*Cyberlindnera jadinii* (*Candida utilis*), *Naganishia* (*Cryptococcus*) *diffluens, Pichia cactophila* (*Candida inconspicua*), *Rhodotorula mucilaginosa, Wickerhamiella* (*Candida*) *pararugosa, Wickerhamomyces anomalus* (*Candida pelliculosa*)) were correctly identified even though they were also underrepresented (<5 isolates per species) in the reference spectra database.

In addition to the 534 isolates included in this prospective evaluation study, an additional evaluation of the performance of the ATR-FTIR-based method for the identification of *C. auris* was undertaken. Since this species was rarely found in routine (*n*=1) during this study period, an additional 24 *C. auris* isolates were solicited from the CDC to test the performance of the ATR-FTIR-based identification method. Employing the reference database, all 24 isolates were correctly identified to the species level with high confidence. In addition, the supplementary 39 randomly selected isolates belonging to 8 species of *Candida* (Table A.10) provided by CHUSJ children's hospital were also all correctly identified to the species level.

Table 6.1. Performance of centrally created ATR-FTIR spectral reference database in a prospective multicenter evaluation study for the identification of clinical yeasts

	No. of isolates (%)				
Microorganism (previous name)	In database	Collected	Correct ID	Misidentified	No identification[1]
Candida albicans	16	242	242	0	0
Candida auris	11	1	1	0	0
Candida dubliniensis	13	25	24 (96)	0	1 (4.00)
Candida (Nakaseomyces) glabrata	17	76	75 (98.7)	0	1 (1.32)
Candida metapsilosis	1	1	0	0	1
Candida orthopsilosis	9	3	2 (66.7)	0	1 (33.3)
Candida parapsilosis	12	79	79	0	0
Candida tropicalis	15	36	36	0	0
Clavispora (*Candida*) *lusitaniae*	10	18	18	0	0
Cryptococcus neoformans	9	2	2	0	0
Cyberlindnera jadinii (*Candida utilis*)	4	1	1	0	0
Kluyveromyces marxianus (*Candida kefyr*)	17	2	2	0	0
Meyerozyma caribbica (*Candida fermentati*)	5	1	1	0	0
Meyerozyma (*Candida*) *guilliermondii*	15	6	4 (66.7)	0	2 (33.3)
Naganishia (*Cryptococcus*) *diffluens*	2	1	1	0	0
Pichia cactophila (*Candida inconspicua*)	1	1	1	0	0
Pichia kudriavzevii (*Candida krusei*)	10	17	17	0	0
Rhodotorula mucilaginosa	3	2	2	0	0
Saccharomyces cerevisiae	13	2	2	0	0
Saprochaete clavata (*Geotrichum clavatum*)	1	1	0	0	1
Starmerella (*Candida*) *magnoliae*	1	1	0	0	1
Torulaspora delbrueckii (*Candida colliculosa*)	0	1	0	0	1
Trichosporon asahii[2]	4	2	2	0	0
Trichosporon (Cutaneotrichosporon) dermatis[2]	1	1	1	0	0
Trichosporon inkin[2]	1	1	1	0	0
Trichosporon (*Apiotrichum*) *mycotoxinivorans*[2]	1	1	1	0	0
Wickerhamiella (*Candida*) *pararugosa*	4	3	3	0	0
Wickerhamomyces anomalus (*Candida pelliculosa*)	2	5	5	0	0
Yarrowia (*Candida*) *lipolytica*	7	2	2	0	0
TOTAL	205	534	525 (98.3)	0	9 (1.7)

[1]No identification is defined by having the triplicate spectra of the single (unknown) isolate predicting as 3 different organisms in the spectral database, and/or having a large standard deviation of the spectral similarity from the mean spectrum (>3.0) of the predicted yeast and/or a spectral similarity of less than 70%.

[2]Species of the genus *Trichosporon* were identified only to the genus level by the ATR-FTIR spectroscopy-based method.

6.5. DISCUSSION

Employing the previously constructed ATR-FTIR spectral database representing 65 yeast species from 23 genera, this prospective multicenter evaluation study resulted in 98.3% correct species identification with no misidentification at the genus level, and 1.7% reported as inconclusive at the species level. While 4 isolates in this multicenter evaluation study were not identified as a result of being underrepresented in the reference database (<5 isolates per species), 5 isolates were not identified while they were represented (>9 isolates per species) in the reference spectral database. The latter finding illustrates the need for augmentation of the spectral database for certain species that likely exhibit a larger spectral variability stemming from a broader range of phenotypic/metabolomic diversity and highlights the limitation of selecting an arbitrary cut-off of the minimum number of isolates that must be included in the database. This can be rationalized in terms of spectral diversity inherit to a given genus and its species within. The spectral similarity/dissimilarity (under standardized culture conditions) reflects the biochemical diversity of the whole organism for a given species. Species that have minimal fluctuation in their biochemical composition under specified growth conditions will be well represented with fewer spectra in the reference spectral database, while those that have a larger fluctuation in their metabolome/biochemical products produced will require a greater number of spectra to be included in the spectral database.

For example, *C. glabrata* is well known for its great intra-species diversity (26) and is most notably known for its reduced susceptibility to fluconazole and increasing reported cases of echinocandin resistance (5, 27). Although *C. glabrata* is subjectively adequately represented in the spectral reference database (*n*=17) and yielded 75/76 (98.7%) correct identification, 1 isolate of *C. glabrata* was not identified due to its lack of similarity (SD of spectral similarity from mean of closest match = 16; spectral similarity to closest match = 30.4%) to the reference *C. glabrata* spectra in the reference spectral database. Through observations, the unidentified *C. glabrata* strain relative to those represented in the reference spectral database, exhibited slower growth (smaller colonies) after incubation for 48 h at 30°C, and through broth microdilution, the isolate in question was found to be resistant (64 µg/mL) to fluconazole, while those in the reference database are all

susceptible-dose dependent to fluconazole (or non-resistant). Both observations may have impacted the spectral profile of the isolate and resulted in no identification due to being relatively phenotypically atypical to those in the reference spectral database.

Further spectral investigation of the unidentified *C. glabrata* with antimicrobial susceptibility testing using the Clinical and Laboratory Standards Institute guidelines M60 (Performance Standards for Antifungal Susceptibility Testing of Yeasts) resulted in the minimum inhibition concentration of 64mg/L to fluconazole and is interpreted as being resistant to fluconazole. With that note, isolates in the reference database are all susceptible-dose dependent to fluconazole (or non-resistant) – suggesting ATR-FTIR spectral information may discriminate between difference in antifungal resistance for *C. glabrata*. *C. glabrata* resistance to azoles such as fluconazole have been linked but not limited to the upregulation and overexpression of the ATP-binding cassette (ABC) transporter genes *CgCDR1* and *CgCDR2* (28). Upregulation of ABC transporters (the predominant protein embedded throughout the cell membrane) results in increased efflux activity while overexpression of the gene increases the presence of the ABC transporters (29). Both mentioned activities may greatly affect the biochemistry (change in cell wall composition and structure as well as intercellular composition) of yeast cells and may be observed in the ATR-FTIR spectra (Figure 6.1). Figure 6.2 illustrates high variance between the inconclusive isolate with the *C. glabrata* where variances are associated with C-OH, C-O-C, C-C and P=O bonds of various biomolecules such as carbohydrates found in the cell wall.

Even with quantifiable success of the multicenter study, few hindrances have been observed throughout the course of the research such as varying results with different isolates cultured onto different culture media. For example, LSPQ uses Difco SAB (5.6 ± 0.2 @ 25°C) while CHUS and HMR uses the modified agar Difco SAB Emmons (pH 6.9 ± 0.2 @ 25°C) where the differences mainly lie in the level of acidity and dextrose content; and, ATR-FTIR spectra of a single strain cultured onto both culture media exhibited significantly distinguishable absorbances in several spectral regions including those associated with carbohydrates (Figure 6.2). Accordingly, high species level misidentification rates were observed for various species when the spectra were not

acquired on the same growth media formulation (30-33). The use of standardized culture media and growth conditions is therefore required to achieve the high specificity and sensitivity reported in this study. Future work can address expanding the reference spectral database of reference isolates grown on different culture media or construction culture-medium specific spectral databases.

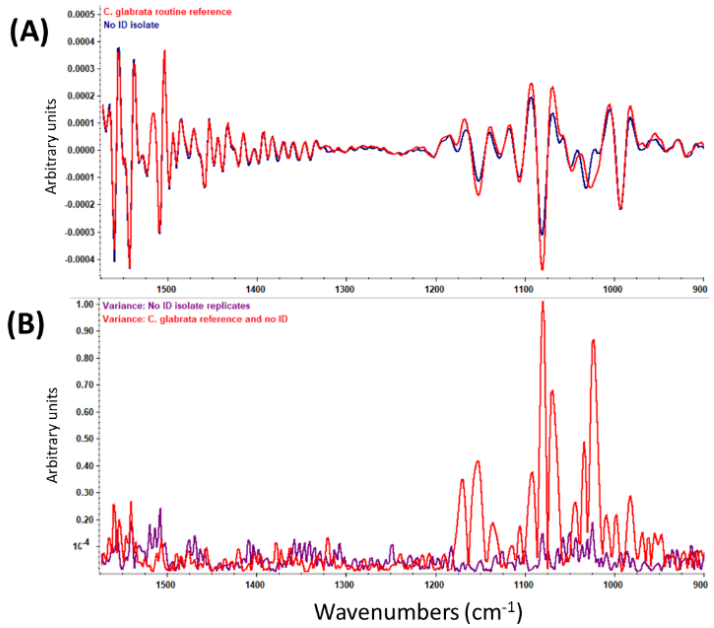


Figure 6.1. (A) Second derivative superimposed spectra belonging to reference *C. glabrata* (red) and the inconclusive identification (blue) sample by the ATR-FTIR spectroscopy-based method (identified as *C. glabrata* by MALDI-TOF MS). (B) represents the variance spectra of the triplicate spectra of the inconclusive result-isolate (purple) and the variance spectrum between the average of the inconclusive result-isolate and reference *C. glabrata* spectrum.

Implementation of the ATR-FTIR-based method must include a highly consistent standard operating procedure, in particular, the use of a consistent growth medium, growth and temperature (14, 34). The use of standardized culture media and growth conditions is therefore required to achieve the high specificity and sensitivity reported in this study. Future work should address expanding the reference spectral database of reference isolates grown on different culture media or the construction of culture-medium specific spectral databases for more robustness and flexibility of this identification method. Without the representation of several culture media, culture conditions, the end user would have to make changes in their routine workflow to conform to the growth media

and conditions utilized in the construction of the reference spectral database. Likewise, ATR-FTIR spectra representative of species diversity can likely be improved by acquiring isolates from varying geographical locations (30, 35, 36).

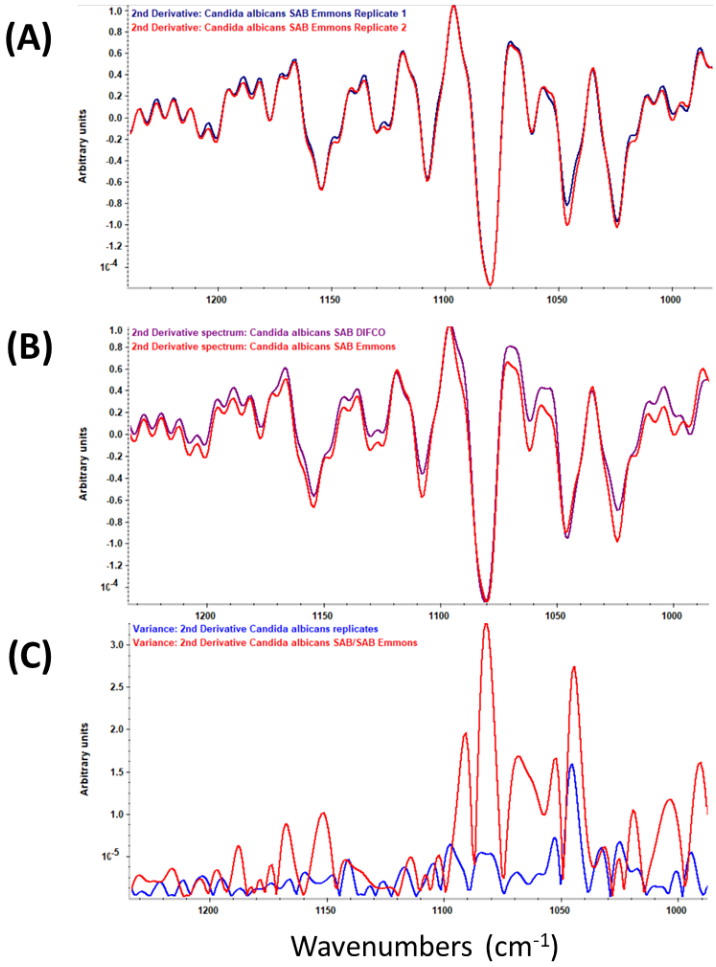


Figure 6.2. (A) Second derivative overlay spectra demonstrating spectral reproducibility of two spectra (replicates) of a strain of *Candida albicans*. (B) Second derivative overlay spectra demonstrating spectral variances between a *Candida albicans* strain grown on Sabouraud dextrose agar (SAB) and modified SAB (SAB Emmons). (C) Variance spectra of second derivative overlay spectra demonstrating higher variance between a strain of *Candida albicans* grown on different agar compared to replicates of the strain

MALDI-TOF MS, which also uses spectral references databases, are constructed with mass spectral profiles of microorganisms representing intra species diversity (i.e., atypical strains, typical strains, isolated from different regions, source of isolation) and varying culturing conditions by having isolates cultured on several culture media and grown in different atmospheric conditions (37).

Lastly, in unpublished work by our group in an exploratory study, the spectral database of yeast developed in Quebec at LSPQ was reconstructed using a different spectrometer (SummitPro, Thermofisher Scientific, MA) was utilized to prospectively identify 41 routine yeasts isolates (belonging to *C. albicans, C. dubliniensis, C. glabrata and C. parapsilosis*) cultured on SAB incubated and 30°C for 24-48 h from an England-based hospital (Maidstone, UK). As a result, 100% correct species identification was achieved. In addition to the success of the multicenter prospective study, the exploratory study has demonstrated the potential applicability of the current spectral database for use globally.

6.6. CONCLUSION

Like MALDI-TOF MS, the ATR-FTIR spectral reference database will need to be constructed to represent species cultured under varying incubation conditions (i.e., time and temperature), use of different culture media, and to include isolates from varying geographical locations, atypical strains, source of isolation, and antimicrobial susceptibility profiles. Additional strains are therefore being solicited from different laboratories to increase species representation in our spectral reference database of yeasts. With 99.0% correct identification and low misidentification rates, our study does however demonstrate the applicability and utility of the ATR-FTIR spectroscopy-based method of reliable species identification of clinically relevant yeasts multicentrally cross province. This is achieved by establishing a well-defined standard operating protocol.

The result from this study delineates further the performance of the first rapid ATR-FTIR-based method for accurate identification of clinical yeasts. The use of a highly standardized ATR-FTIR-based method has been evaluated in a multicentre study demonstrating the reliability of ATR-FTIR spectroscopy for species-level identification of clinically relevant yeasts. This is the first study to demonstrate the robustness of the constructed reference database to achieve >98% correct identification using multiple ATR-FTIR spectrometers at 6 distinct clinical microbiology laboratories and with multiple system operators. Accordingly, the ATR-FTIR based method should be considered for further validation and diagnostic method accreditation as it offers a new reagent-free, cost-effective method and provides species identification within minutes after initial culture. Additionally, it is highly affordable and may be used as an alternative (without

compromising accuracy of results) to costly molecular techniques and, low-accuracy and difficult-to-identify-rare-yeasts biochemical techniques for small and mid-size laboratories. The next step is to incorporate more clinical organisms such as bacteria, mycobacteria and molds to the reference spectral database and the pursuit of the ATR-FTIR-based method accreditation as an *in vitro* diagnostic device facilitating its acceptance and implementation in a clinical setting.

6.7. REFERENCES

1. Bongomin F, Gago S, Oladele RO, Denning DW. 2017. Global and multi-national prevalence of fungal diseases—estimate precision. Journal of fungi 3:57.
2. Brown GD, Denning DW, Gow NA, Levitz SM, Netea MG, White TC. 2012. Hidden killers: human fungal infections. Science translational medicine 4:165rv13-165rv13.
3. Rajasingham R, Smith RM, Park BJ, Jarvis JN, Govender NP, Chiller TM, Denning DW, Loyse A, Boulware DR. 2017. Global burden of disease of HIV-associated cryptococcal meningitis: an updated analysis. The Lancet infectious diseases 17:873-881.
4. Richardson M, Lass-Flörl C. 2008. Changing epidemiology of systemic fungal infections. Clinical Microbiology and Infection 14:5-24.
5. Lamoth F, Lockhart SR, Berkow EL, Calandra T. 2018. Changes in the epidemiological landscape of invasive candidiasis. Journal of Antimicrobial Chemotherapy 73:i4-i13.
6. Pappas PG, Lionakis MS, Arendrup MC, Ostrosky-Zeichner L, Kullberg BJ. 2018. Invasive candidiasis. Nature Reviews Disease Primers 4:1-20.
7. Castanheira M, Messer SA, Rhomberg PR, Pfaller MA. 2016. Antifungal susceptibility patterns of a global collection of fungal isolates: results of the SENTRY Antifungal Surveillance Program (2013). Diagnostic microbiology and infectious disease 85:200-204.
8. Chowdhary A, Sharma C, Meis JF. 2017. Candida auris: a rapidly emerging cause of hospital-acquired multidrug-resistant fungal infections globally. PLoS pathogens 13:e1006290.
9. Gautam V, Sharma M, Singhal L, Kumar S, Kaur P, Tiwari R, Ray P. 2017. MALDI-TOF mass spectrometry: An emerging tool for unequivocal identification of non-fermenting Gram-negative bacilli. The Indian journal of medical research 145:665.

10. Stefaniuk E, Baraniak A, Fortuna M, Hryniewicz W. 2016. Usefulness of CHROMagar Candida medium, biochemical methods–API ID32C and VITEK 2 compact and two MALDI-TOF MS systems for Candida spp. identification. Pol J Microbiol 65:111-4.

11. Murray PR. 2012. What is new in clinical microbiology—microbial identification by MALDI-TOF mass spectrometry: a paper from the 2011 William Beaumont Hospital Symposium on molecular pathology. The Journal of Molecular Diagnostics 14:419-423.

12. Jeffery-Smith A, Taori SK, Schelenz S, Jeffery K, Johnson EM, Borman A, Manuel R, Brown CS. 2018. Candida auris: a review of the literature. Clinical microbiology reviews 31:e00029-17.

13. Seng P, Rolain J-M, Fournier PE, La Scola B, Drancourt M, Raoult D. 2010. MALDI-TOF-mass spectrometry applications in clinical microbiology. Future microbiology 5:1733-1754.

14. Lam LM, Dufresne PJ, Longtin J, Sedman J, Ismail AA. 2019. Reagent-Free Identification of Clinical Yeasts by Use of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy. Journal of clinical microbiology 57:e01739-18.

15. Quilès F, Accoceberry I, Couzigou C, Francius G, Noël T, El-Kirat-Chatel S. 2017. AFM combined to ATR-FTIR reveals Candida cell wall changes under caspofungin treatment. Nanoscale 9:13731-13738.

16. Costa FS, Silva PP, Morais CL, Arantes TD, Milan EP, Theodoro RC, Lima KM. 2016. Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii. Analytical Methods 8:7107-7115.

17. Silva S, Tobaldini-Valerio F, Costa-de-Oliveira S, Henriques M, Azeredo J, Ferreira EC, Lopes JA, Sousa C. 2016. Discrimination of clinically relevant Candida species by Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR-ATR). RSC Advances 6:92065-92072.

18. Essendoubi M, Toubas D, Lepouse C, Leon A, Bourgeade F, Pinon J-M, Manfait M, Sockalingum GD. 2007. Epidemiological investigation and typing of Candida glabrata clinical isolates by FTIR spectroscopy. Journal of microbiological methods 71:325-331.

19. Toubas D, Essendoubi M, Adt I, Pinon J-M, Manfait M, Sockalingum GD. 2007. FTIR spectroscopy in medical mycology: applications to the differentiation and typing of Candida. Analytical and bioanalytical chemistry 387:1729.

20. Lecellier A, Gaydou V, Mounier J, Hermet A, Castrec L, Barbier G, Ablain W, Manfait M, Toubas D, Sockalingum G. 2015. Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds. Food microbiology 45:126-134.

21. Taha M, Hassan M, Essa S, Tartor Y. 2013. Use of Fourier transform infrared spectroscopy (FTIR) spectroscopy for rapid and accurate identification of yeasts isolated from human and animals. International journal of veterinary science and medicine 1:15-20.

22. Naumann A. 2015. Fourier transform infrared (FTIR) microscopy and imaging of fungi, p 61-88, Advanced Microscopy in Mycology. Springer.

23. Novais Â, Freitas AR, Rodrigues C, Peixe L. 2019. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. European Journal of Clinical Microbiology & Infectious Diseases 38:427-448.

24. Kurtzman C, Robnett C. 1997. Identification of clinically important ascomycetous yeasts based on nucleotide divergence in the 5'end of the large-subunit (26S) ribosomal DNA gene. Journal of clinical microbiology 35:1216-1223.

25. White TJ, Bruns T, Lee S, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols: a guide to methods and applications 18:315-322.

26. Carreté L, Ksiezopolska E, Gómez-Molero E, Angoulvant A, Bader O, Fairhead C, Gabaldón T. 2019. Genome Comparisons of Candida glabrata Serial Clinical

Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations. Frontiers in Microbiology 10.

27. Howell SA, Hazen KC, Brandt ME. 2015. Candida, Cryptococcus, and other yeasts of medical importance. Manual of clinical microbiology:1984-2014.

28. Sanglard D, Ischer F, Bille J. 2001. Role of ATP-binding-cassette transporter genes in high-frequency acquisition of resistance to azole antifungals in Candida glabrata. Antimicrobial Agents and Chemotherapy 45:1174-1183.

29. Prasad R, Goffeau A. 2012. Yeast ATP-binding cassette transporters conferring multidrug resistance. Annual review of microbiology 66:39-63.

30. Carranza L. 2012. Standardization and internal validation of a bacteria identification method utilizing focal-plane-array Fourier transformed infrared spectroscopy. McGill University (Canada).

31. Haag H, Gremlich H-U, Bergmann R, Sanglier J-J. 1996. Characterization and identification of actinomycetes by FT-IR spectroscopy. Journal of microbiological methods 27:157-163.

32. Lefier D, Hirst D, Holt C, Williams AG. 1997. Effect of sampling procedure and strain variation in Listeria monocytogenes on the discrimination of species in the genus Listeria by Fourier transform infrared spectroscopy and canonical variates analysis. FEMS microbiology letters 147:45-50.

33. Wenning M, Breitenwieser F, Konrad R, Huber I, Busch U, Scherer S. 2014. Identification and differentiation of food-related bacteria: a comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry. Journal of microbiological methods 103:44-52.

34. Pebotuwa S, Kochan K, Peleg A, Wood BR, Heraud P. 2020. Influence of the Sample Preparation Method in Discriminating Candida spp. Using ATR-FTIR Spectroscopy. Molecules 25:1551.

35. Wehrli PM, Lindberg E, Svensson O, Sparén A, Josefson M, Dunstan RH, Wold AE, Gottfries J. 2014. Exploring bacterial phenotypic diversity using factorial

design and FTIR multivariate fingerprinting. Journal of Chemometrics 28:S681-S686.

36. Ribeiro da Cunha B, Fonseca LP, Calado CR. 2019. A phenotypic screening bioassay for Escherichia coli stress and antibiotic responses based on Fourier-transform infrared (FTIR) spectroscopy and multivariate analysis. Journal of applied microbiology 127:1776-1789.

37. biomerieux-diagnostics. 2020. VITEK® MS Mass spectrometry microbial identification system: Comprehensive, regularly-updated database. https://www.biomerieux-diagnostics.com/vitekr-ms-0#VITEK%C2%AE%C2%A0SOLUTIONS:%20Complete%20traceability%20and%20flexibility. Accessed

Connecting statement 5

ATR-FTIR spectroscopy in microbiology may be applicable for investigating phenotypic attributes of microorganisms and/or may be a useful tool for classifying microorganisms and it can also achieve identification to genotypic methods if the gene variability can be associated with unique phenotypic attributes. As such, the following chapter will investigate phenotypic and genotypic characteristics of a potential *Candida parapsilosis* outbreak in Chile relative to ATR-FTIR spectral data and evaluate the method for epidemiological strain-typing for potential outbreak detection.

CHAPTER 7.CASE STUDY: RETROSPECTIVE OUTBREAK INVESTIGATION AND COMPARISON OF *CANDIDA PARAPSILOSIS SENSU STRICTO* OBTAINED IN CHILE BY MOLECULAR SPECTROSCOPY AND PHENOTYPIC TECHNIQUES

7.1. ABSTRACT

Candida is the fourth leading causative agent of hospital-acquired bloodstream infections and is associated with high mortality rates. Among *Candida* spp., *Candida parapsilosis* is frequently implicated in outbreaks owing to its ability to form biofilms, binding to and surviving on venous catheters. Current epidemiological methods such as multi-locus sequence typing, microsatellite molecular typing (MSMT), and whole genome sequencing (WGS) are labor-intensive, require highly trained personnel, and are costly, making it impractical to employ these methods for routine typing of suspected outbreak isolates. Previous studies performed by our group have successfully employed attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy as a non-destructive, rapid, and reagent-free technique for yeast identification. In the present study, the potential application of this inexpensive whole-organism fingerprinting technique as a reagent-free typing method was evaluated with a set of 41 *C. parapsilosis* isolates obtained from 17 hemodialysis clinics belonging to 9 Chilean hospitals from suspected outbreaks in 2012 and 2013. MSMT of *C. parapsilosis* was achieved by selecting 5 polymorphic microsatellite markers (B, G, CP1α, CP4α and CP6α). Additionally, a subset of the isolates was selected for WGS and ^{1}H high resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy to compare to ATR-FTIR spectroscopy and MSMT results. As a result, MSMT of the 41 *C. parapsilosis* isolates resulted in 6 genotypes encompassing >2 isolates sharing the same microsatellite profiles (clusters identified as CL1 through CL6) – suggesting a multicenter polyclonal outbreak. CL1-CL2 and CL3-CL4 are closely related (with a difference of one and two alleles, respectively), with CL5 and CL6 being the most dissimilar from the others. ATR-FTIR and HR-MAS NMR spectroscopy and WGS were in agreeance with for 3 of the 2 evaluated MSMT clusters. This study demonstrates that ATR-FTIR spectroscopy provides comparable discriminatory power to MSMT and WGS for outbreak detection, however, additional research for additional cases will make the technique a potentially a

useful tool for prospective outbreak surveillance. NMR spectroscopy may also aid in biomarker discovery.

7.2. INTRODUCTION

While *Candida albicans* is the leading cause of candidemia worldwide [high mortality rate of ~40% in developed countries (1)], the prevalence of non-*albicans* species such as *Candida glabrata* have risen in Northern Europe and the United States of America in the recent decades. In Latin America, Southern Europe and Asia, *Candida parapsilosis* and *Candida tropicalis* are the leading cause of non-*albicans* species candidemia (2-4). *C. parapsilosis* is a natural skin colonizer and is one of the most frequently isolated yeast on the hands of healthcare workers and is found ubiquitously in nature (animals, aquatic environment, soils and insects) (5). The transmission of *C. parapsilosis* from healthcare worker-to-patient is often from the installation and maintenance of intravascular catheters commonly effecting neonates and surgical intensive care unit patients (6). Due to its capacity to bind and survive on catheters and other indwelling plastic medical devices, *C. parapsilosis* to is a common causative agent of bloodstream hospital-acquired infections. This microorganism is also well known for its capacity to form biofilms on medical implants and once embedded, *C. parapsilosis* is particularly resistant to antifungal treatment (7). Moreover, *C. parapsilosis* is horizontally transmitted and therefore can easily cause outbreaks and several centers have reported an increase in invasive *C. parapsilosis* infections (1, 3, 8). One study determined that 72% of 1240 isolates collected were clonal (3), suggesting the need for rigorous infection control and rapid strain identification techniques. For example, although China has few cases of *C. parapsilosis* outbreaks, 17 candidemia cases were observed in the neonate intensive care unit in Beijing where 2 cases were fatal. After random amplified polymorphic DNA (RAPD) typing, all isolates were identical and infection control measures were implemented to successfully control the outbreak (9). Rapid strain-identification is therefore needed to control outbreaks in hospitals.

Currently, molecular genotyping methods [RAPD, microsatellite genotyping and DNA sequencing of the internal transcribed spacer (ITS) and D1/D2 regions of the 18S and 28S rRNA gene by polymerase chain reaction (PCR)] are successfully employed to strain

type *C. parapsilosis* (3, 9-13). Although molecular methods achieve higher identification and strain-typing sensitivity compared to serologic and phenotypic methods which are considered to be the gold standards for identifying invasive fungal microorganisms (14), they are not financially feasible for resource limited laboratories, are time consuming for routine diagnostics and require expertise to perform the assays (15).

Interestingly, in 2009, the complete WGS of *C. parapsilosis* was achieved (4). Although WGS has gained popularity and is becoming more affordable for research laboratories, WGS is still not an alternative for routine use or resources limited laboratories and requires highly trained personnel for data acquisition and analysis. In addition, WGS may not be the solution to strain-type certain species for outbreak investigation as some species such as *C. albicans* display genetic plasticity under different types of stresses (i.e., presence of antifungal, heat shock, growth in host) (16). Moreover, WGS of emerging multidrug resistance *Candida auris* results in low genetic diversity between strain for the same geographical clade (i.e., Africa, East Asia, South Asia and South America) with a difference of less than 60 single nucleotide polymorphism (SNP) and tens of thousands of SNPs differences between geographical clades (17). Unlike *C. auris*, WGS of *C. parapsilosis* shows low genetic variation between strains of different geographical origins such as Europe and North America, while major genetic variations are due to genes that are responsible for differences in cell surface glycoproteins associated with host-pathogen interactions (4). With varying strain-typing results for different species, it is suggested that WGS for some yeast species may be more useful to analyze antifungal resistance profiles rather than for epidemiology of nosocomial outbreaks (18).

The applicability of Fourier transform infrared (FTIR) spectroscopy for microbial identification has gained momentum over the past decade (19-25). Even with numerous successful studies, there is a lack of robustness of the models developed for implementation for routine use. Unlike molecular genotypic methods which require analysis of RNA/DNA sequences to those in a reference database or by the sequence similarity relative to each other, the infrared spectrum of a microorganism (whole-organism fingerprint) can be acquired directly from intact cells taken from a culture plate

and is representative of their complete biochemical composition (i.e., lipids, proteins, carbohydrates, etc.) (26). In addition to successfully identifying a wide range of species, FTIR spectroscopy has also been widely studied for strain-typing (i.e., patient-to-patient, outbreak investigation, source of isolation) (27-36). The use of attenuated total reflectance (ATR) mode for acquiring FTIR spectra makes it a rapid, reagent-free and affordable alternative to existing microbial identification and strain-typing techniques.

Similar to ATR-FTIR spectroscopy, high-resolution magic-angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy is a non-targeted whole organism fingerprinting technique. In microbiology, most commonly, NMR spectroscopy is used to observe *in vivo* metabolic profiles of microorganisms in the presence of stress factors such as antimicrobials, metabolite discoveries and cell wall structural characterization (37-41). HR-MAS NMR spectra collected from intact microorganism illustrates the complete biochemical components (i.e., lipids, proteins, carbohydrates, etc.) that are associated with the observed nuclei (i.e., ^{1}H, ^{13}C, ^{31}P). Various studies have evaluated HR-MAS NMR spectroscopy for microbial identification and strain typing with successful results compared to conventional methods both in our laboratory and others (42-44). Both ATR-FTIR and HR-MAS NMR spectroscopy are non-destructive and reagent-free techniques with high discriminatory power. Comparing NMR and FTIR spectroscopy for differentiating phylogenetically similar yeasts such as *C. albicans* and *Candida dubliniensis* provided similar results (42, 45). Combining information collected from both HR-MAS and ATR-FTIR spectroscopy may allow for high discrimination power of species and potentially aid in strain typing for the investigation of outbreaks and biomarker discovery.

In this study, Chilean outbreak strains of *C. parapsilosis sensu stricto* were collected from Chile's public health reference laboratory and analyzed and compared by microsatellite genotyping, WGS and, ATR-FTIR and HR-MAS NMR spectroscopy. Evaluation of the spectroscopic methods with the currently accepted microsatellite genotyping techniques will determine the most sensitive and rapid strain-typing method for *C. parapsilosis* complex and *C. parapsilosis sensu stricto,* allowing for early detection, rapid identification thereby reducing mortality rates during outbreak and epidemiological investigations.

7.3. MATERIALS AND METHODS

7.3.1. Sample collection

A total of 42 *in situ C. parapsilosis* (Table A.10) were retrospectively collected from Chilean hospitals of a suspected outbreak and sent to Quebec's microbiology reference laboratory, Laboratoire de Santé Publique du Québec (LSPQ) for the evaluation. The isolates were collected from 9 hospitals and 17 dialysis centers from frozen cultures between 2012-2013 from infected patients or probable carriers (Table A.10).

7.3.2. Purity assessment

To assess for the presence of mixed isolates, all specimens were re-streaked onto Sabouraud dextrose agar (SAB; BBL™) and BBL™ CHROMagar™ *Candida* (Becton Dickinson, Sparks, MD, USA) plates. Any heterogenous colonies-looking plates were re-streaked onto both SAB and CHROMagar™ *Candida* from an isolated colony.

7.3.3. Antimicrobial susceptibility testing

The selected 10 isolates (Table 7.1) that were used for WGS, ATR-FTIR spectroscopy, NMR spectroscopy, microsatellite molecular strain typing (MSMT), and antimicrobial susceptibility testing (AST) was achieved by following the Clinical and Laboratory Standards Institute (CLSI) reference method for broth dilution antifungal susceptibility testing of yeasts (M27Ed4). In brief, antifungals such as amphotericin B, flucytosine, echinocandins and azoles of varying concentrations (0.008-128 µg/mL) were dehydrated into a 96-well microtiter plate suspended in 100 µL specialty microbiology RMPI-1640 broth as indicated in

Table A.12. Per each sample that was cultured onto SAB and incubated for 24 h at 35°C, 0.5 MacFarland standards in 0.85% saline and confirming a 0.76% transmittance using a spectrophotometer at 530 nm to create a 1-5x10^6 cells/ mL suspension. The suspension is diluted in RMPI-1640 culture medium to create a final 0.5x10^3 to 2.5x10^3 CFU/mL solution. In the 96-well plate (Table A.12), 100 µL of the suspension was inoculated into each well and incubated for 24 h at 35°C.

Table 7.1. Selected details of isolates of varying microsatellite clusters for strain typing analysis by ATR-FTIR spectroscopy, HR-MAS NMR spectroscopy and WGS

LSPQ ID2	MST Cluster	Hospital	Center	Date of isolation	Halotype
MY076475	CL1	Aguirre	Aguirre	04-11-2012	AB[AC]E[AC]
MY077970	CL1	Hurtado	San Gabriel	02-20-2013	AB[AC]E[AC]
MY077968	CL2	Hurtado	CD4	12-07-2012	AB[AC]EC
MY077962	CL2	Pto Montt	Pto Montt	04-30-2013	AB[AC]EC
MY077956	CL3	Hurtado	CD3	12-06-2012	ABA[AB]A
MY077961	CL3	Hurtado	La Serena	01-25-2013	ABA[AB]A
MY077958	CL4	Hurtado	Rosita	03-05-2013	ABAB[BC]
MY077959	CL4	San Borja	San Borja	01-25-2013	ABAB[BC]
MY077952	CL5	San Borja	San Borja	11-08-2012	BA[AB]C[DE]
MY077954	CL5	San Borja	San Borja	03-09-2013	BA[AB]C[DE]

7.3.4. Microscopic assessment

Of the 10 isolates used for WGS, MSMT, AST, ATR-FTIR spectroscopy, and NMR spectroscopy, the isolates were also grown on cornmeal agar with tween 80. The agar was scratched with a sample-inoculated pick where the surface of the scratched agar was covered with a sterile cover slip and incubated for 48 h at 30°C. Microscopic images were collected using an Olympus microscope with the 10x and 40x magnification to capture the cellular morphology of the samples grown on the plate.

7.3.5. Species identification

7.3.5.1. Ribosomal DNA sequencing for confirmation of *C. parapsilosis* designation.

All strains were PCR amplified and sequenced at the commonly used loci for *C. parapsilosis* complex species discrimination (D1/D2 and ITS regions at the 5' end of the 18S rDNA). Identification was obtained by BLAST similarity searches against those found GenBank nr nucleotide database at the National Center for Biotechnology Information and that of CBS-KNAW Fungal Biodiversity Centre.

7.3.5.2. MALDI-TOF MS species identification

All isolates were identified by MALDI-TOF MS using manufacturer guidelines. In brief, after a *C. parapsilosis* isolates are grown on a culture media agar, a sample is deposited and treated with 0.5 µL formic acid onto a target plate, air-dried then treated with 1 µL α-

cyano-4-hydroxycinnamic acid, air-dried and inserted into the MALDI-TOF MS system which is placed under vacuum. A nitrogen laser is utilized to vaporize the biomolecules within the microorganisms to yield a mass profile of the biomolecules unique to the given species to which the microorganism belongs. The MALDI-TOF mass spectrum is then compared to other spectra in the commercial MALDI-TOF MS spectral database to confirm the identity of the isolate as *C. parapsilosis*.

7.3.5.3. ATR-FTIR spectroscopy-based species identification

In addition to rDNA sequencing for species confirmation of the isolates, all strains where interrogated in a previously created ATR-FTIR spectral database (19). Sample preparation and spectral acquisition procedure follow the one previously described in the thesis. In brief, colonies were picked directly from 48 h incubated plates (at 30°C) and deposited onto the ATR sampling surface per sample in triplicate. Sixty-four background co-added scans were acquired before every 64 co-added sample scans taken at 8 cm^{-1} spectral resolution.

7.3.6. Strain typing techniques

7.3.6.1. ATR-FTIR spectroscopy strain-typing method

Spectra collected using the methodology described in section 7.3.5.3 were assigned individual class designation and subjected to a feature selection algorithm to determine the spectral regions associated with the highest discrimination between all “classes”, with the assumption that each individual *C. parapsilosis* isolate is different from the others. The triplicate spectra (per sample) were then averaged and analyzed by hierarchical cluster analysis (HCA) to determine relative spectral similarity distances between the observed replicates and the different classes.

7.3.6.2. NMR spectroscopy

NMR spectra were collected on a Bruker Ascend™ 600 MHz NMR spectrometer. The microorganisms were cultured using the same methods for MALDI-TOF MS and ATR-FTIR spectroscopy-based techniques for microorganism identification. Colonies on the culture plate were harvested with a sterile needle and transferred into and NMR insert,

sealed, and placed into a rotor containing an NMR spectroscopy reference compound (trimethylsilylpropanoic acid) dissolved in D_2O. Spectra were collected using a ^1H probe, the ^1H NMR acquisition parameter are as follows: frequency 600 MHz, pulse angle 54.7º and centered a 0 ppm in the 0-6 ppm spectral region. One spectrum was acquired per isolate (10 selected) where they were analyzed by HCA to determine relative spectral similarity distances between the observed 10 isolates.

7.3.6.3. Microsatellite genotyping

Five polymorphic microsatellite markers were selected: markers B, G, CP1α, CP4α and CP6α. These markers have successfully been used for molecular typing of *C. parapsilosis* strains (46, 47). Each of these markers can be mapped to a distinct genetic locus found at least 200 kilobases from one another. The markers B, G and CP6α are found, likely, on the same chromosome (46, 47).

Microsatellite PCR reactions of 50 µL were composed of: 5X Phusion HF buffer (1X), 200 nM of each primer, 200 nM de dNTP, 1 U of Phusion HF DNA polymerase and 2 µL of *C. parapsilosis* DNA prepared according to PR-MY-038 method. The following PCR cycle was used: 95°C for 3 min followed by 35 amplification cycles (95°C 30 sec, 58°C 30 seconds and 72°C for 1 min). The amplified fragments were then separated on a 3% agarose gel (2% Nusieve GTG and 1% Seakem GTG). The amplified DNA bands were visualized under UV light after ethidium bromide staining (Figure A.8) (PCR primers used data is not shown).

7.3.6.4. Whole genome sequencing

Gene extraction was achieved by preparing a 2 McFarland sample with glass beads in a 2 ml Eppendorf tube with 1 mL buffer solution. DNA was automatically extracted with NucliSense EasyMag 2.1 (BioMérieux, Marcy L'Étoile). Amplification of the DNA is achieved by PCR and pulsed field gel electrophoresis to validate the DNA have been extracted. WGS is completed by Illumina and cleaning of reads was achieved with Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) and quality control (before and after cleaning of reads) of the reads was completed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Epidemiological analysis

[core single nucleotide variants] was achieved by mapping of the reads, variant calling, to a reference *C. parapsilosis* strain CDC317 (957321 bp genomic sequence, reference in NCBI: HE605209.1) and construction of a pseudo-alignment with SNVPhyl (https://snvphyl.readthedocs.io/en/v1.0/install/versions/).

7.4. RESULTS AND DISCUSSION

7.4.1. Purity assessment

All the colonies, apart from two, displayed a uniform purple-pink color on CHROMagar™ *Candida* media and smooth white color on SAB; consistent with a pure culture of *C. parapsilosis.* The two strains that displayed mixed coloration from cream to purple-pink on CHROMagar™ *Candida* were re-streaked (single colony) onto another CHROMagar™ *Candida* plate and mixed coloration was observed when an isolated colony from both strains was streaked once again on CHROMagar™ *Candida*, concluding that these isolates are pure but naturally show differential coloration on CHROMagar™ *Candida*.

7.4.2. Species identification

C. parapsilosis (*sensu stricto*) can possibly be mistaken for *Candida orthopsilosis* and *Candida metapsilosis*. Both of these species were formerly classified under *C. parapsilosis* groups II and III (48) and are considered part of *C. parapsilosis* complex. To confirm that the isolates received from Chile Public Health Institute were bona fide *C. parapsilosis* strains, the D1/D2 and ITS regions of the ribosomal DNA were sequenced. The DNA sequences of 41 strains were found to be the same and bear 100% identity to *C. parapsilosis* nucleotide records found listed on both the NCBI Genbank (e.g., AB741060.1 Genbank record) and CBS-KNAW Fungal Biodiversity Centre databases. However, one isolate was identified as *C. orthopsilosis* (strain MY077963). MALDI-TOF MS, WGS and ATR-FTIR spectroscopy method of microorganism identification all were in agreeance with the PCR gene sequencing identification. The one strain of *C. orthopsilosis* collected was omitted from the study.

Interestingly, the *C. parapsilosis* were correctly identified by the ATR-FTIR spectroscopy-based method. The current ATR-FTIR spectral database only consists of Canadian yeasts and a handful of yeasts from the United Kingdom, however, all the Chilean isolates were correctly identified, even though *C. parapsilosis* has genetic plasticity and varies geographically, but the genetic differences did not hamper its identification by ATR-FTIR spectroscopy.

7.4.3. Microsatellite molecular typing

MSMT of the 41 *C. parapsilosis* isolates resulted in 6 genotypes encompassing >2 isolates sharing the same microsatellite profiles (clusters identified as CL1 through CL6 and are color coordinated throughout the current chapter) – suggesting a multicenter polyclonal outbreak (Table 7.2). CL1-CL2 and CL3-CL4 are closely related (with a difference of one and two alleles, respectively), with CL5 and CL6 being the most dissimilar from the others (Table 7.2). Details per isolate can be found in Table A.10. Of the 41 isolates, 26 isolates were part of the potential polyclonal outbreak while the other isolates exhibited halotypes that were unique (one of a kind).

Table 7.2. MSMT clusters assignments based on selected microsatellite markers and alleles determined by gel electrophoresis with ATR-FTIR cluster assignment relative to MSMT clusters.

CLUSTERS	MICROSATELLITE MARKERS AND ALLELES				
MSMT clusters	B	G	CP1α	CP4α	CP6α
CL1	A	B	AC	E	AC
CL2	A	B	AC	E	C
CL3	A	B	A	AB	A
CL4	A	B	A	B	BC
CL5	B	A	AB	C	DE
CL6	D	A	A	F	C

The study is focused on strain typing between the MSMT clusters and how other strain typing techniques correlate with the standard method for strain typing. While there are 6 clusters of potential outbreak strains, CL6 through further investigation consists of 2 isolates which were obtained from the same hospital and same patient 19 days apart. Since the isolate was isolated from the same patient, it is not of interest for comparison with the other strain typing methods.

As such, due to limited resources, only 10 isolates of the 24 isolates belonging to MSMT clusters CL1-CL5 were examined by WGS, AST, ATR-FTIR spectroscopy, and NMR spectroscopy. Two isolates of each MSMT cluster were selected and details of the isolates can be found in Table 7.1.

7.4.4. Antibiogram phenotypes of selected isolates

None of the 10 isolates selected had the same antimicrobial susceptibility profiles. Isolate MY076475 (CL1) and MY077957 (CL2) has a difference of 1 reading of minimum inhibition concentration (MIC) of caspofungin and isolate MY077970 (CL1) and MY077962 (CL2) also has a difference of 1 reading of MIC, also for the drug caspofungin. Additionally, the 2 isolates from MSMT CL3 also have differences of 1 reading, but for a different antifungal, micafungin.

Table 7.3. Minimal inhibitory concentrations of antifungals in µg/mL of 10 *Candida parapsilosis* isolates and 2 reference isolates and microsatellite strain type (MSMT)

	MSMT	**Minimal inhibitory concentrations of antifungals in µg/mL**								
Isolate #	**Cluster**	**5FC[1]**	**AMP[2]**	**ANI[3]**	**CAS[4]**	**MIC[5]**	**FLU[6]**	**ITR[7]**	**POS[8]**	**VOR[9]**
LSPQ-01609	REF	8	1	0.06	0.12	0.12	16	0.12	0.03	0.12
LSPQ-01610	REF	2	0.25	1	0.25	0.5	4	0.12	0.06	0.06
MY076475	CL1	2	0.25	1	0.5	1	0.5	0.03	0.016	0.016
MY077952	CL5	2	0.25	1	0.25	0.5	2	0.12	0.03	0.03
MY077954	CL5	2	0.05	2	0.25	1	2	0.12	0.06	0.06
MY077956	CL3	2	0.5	2	0.5	2	0.5	0.06	0.016	0.016
MY077957	CL2	2	0.25	1	1	1	0.5	0.03	0.016	0.016
MY077958	CL4	2	0.25	2	0.25	0.5	0.5	0.03	0.016	0.016
MY077959	CL4	2	0.5	2	0.25	1	0.5	0.016	0.016	0.016
MY077961	CL3	2	0.5	2	0.5	1	0.5	0.06	0.016	0.016
MY077962	CL2	2	0.12	1	1	1	0.5	0.016	0.016	0.016
MY077970	CL1	2	0.12	1	0.5	1	0.5	0.016	0.016	0.016

[1]Flucytosine, [2]Amphotericin B, [3]Anidulafungin, [4]Caspofungin, [5]Micafungin, [6]Fluconazole, [7]Itraconazole, [8]Posaconazole, [9]Voriconazole

Although through MSMT and achieving like-halotypes of at least 2 isolates, their antifungal susceptibility profiles are different. As such, AST for strain-typing is not a reliable method to ascertain isolates are of the same strain type. Phenotypic characteristic such as the MIC of various antifungal varies with the environment and several stress

factors and therefore are actively changing. Although not a good method for strain-typing, it may be useful for determining clustering patterns based on spectrotyping techniques to be discussed in the following sections.

7.4.5. Microscopic phenotype of selected isolates

Like the antibiogram of the 10 selected potential outbreak strains of *C. parapsilosis*, additional phenotypic analysis such as the cell morphology were analyzed to provide additional information to aid in describing relationships of the strains as a prelude to spectroscopic analysis. As a result of observing the strains under the microscope at 10x and 40x magnification, 6 different morphologies were observed (Figure A.9). Table 7.4 indicates which cell morphology was observed for the 10 selected isolates relative to the MSMT clusters.

Table 7.4. Classification of potential Chilean *Candida parapsilosis* isolates by cell morphology grown on corn meal agar

Isolate	MSMT cluster	Cell morphology classification
MY076475	CL1	CM1
MY077952	CL5	CM2
MY077954	CL5	CM3
MY077956	CL3	CM4
MY077957	CL2	CM5
MY077958	CL4	CM3
MY077959	CL4	CM3
MY077961	CL3	CM5
MY077962	CL2	CM6
MY077970	CL1	CM3

The cell morphology of both isolates belonging to MSMT CL4 were observed to have the same cell type/growth, however, two other isolates not belonging to the CL4 clusters were also demonstrated the same cell type. As such, for obvious reasons, cell morphology is an inefficient tool for strain typing, but is traditionally standard for mycologists to identify species through microscopy.

7.4.6. Comparison of alternative strain typing techniques methods to microsatellite molecular strain typing

7.4.6.1. Whole genome sequencing

Though MSMT is widely used to strain type *C. parapsilosis*, it is not a standardized method for epidemiological studies, WGS is much more specific than MSMT and standard for surveillance but has yet to be implemented due to high cost, requiring experts, long result turn around time (yeasts has genome that is approximately 37 Mbp, bacteria is approximately 3.7 Mbp for comparison) and gap in research knowledge on gene function of multiple genes. Clusters generated from WGS differences of the 10 selected *C. parapsilosis* isolates were compared to clusters obtained from MSMT. WGS obtained 4 clusters as opposed to 5 clusters by MSMT (Figure 7.1). WGS clusters are labeled as W1, W2, W3 and W4.

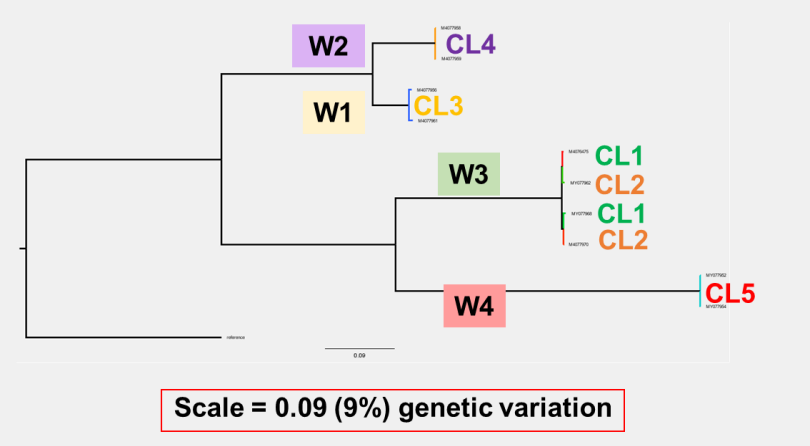


Figure 7.1. Core single nucleotide variant whole genome sequencing (WGS) analysis illustrated in a Newick phylogenetic tree of selected *C. parapsilosis* outbreak strains – colors resemble microsatellite molecular strain typing clusters in the current chapter. WGS clusters are labeled as W1, W2, W3 and W4

MSMT clusters CL1 and CL2 were grouped into one cluster with <1% genetic variation and 13-34 SNPs difference (W3), suggesting clonality between the two MSMT clusters (Table A.14). Moreover, the isolate in CL5 (W4) by WGS were clustered together with a single SNP difference and those isolates in CL4 (W2) also clustered together but with a SNP difference of 4. Lastly, isolates of CL3 (W1) also were in agreeance with MSMT and clustered together with a SNP difference of 31 (Table A.14). As such, based on WGS results, CL3 and CL4 are least similar to the other MSMT cluster (~81% genetic variation),

CL1 and CL2 are the same strains, and CL3 is more similar to CL1-CL2 than CL3-CL4 (this pair has ~14% genetic variation) (Figure 7.1).

7.4.6.2. ATR-FTIR spectroscopy

ATR-FTIR spectra was acquired from all 41 isolates. Through the spectrotyping technique described in the previous chapters, each ATR-FTIR was labelled based on MSMT classification. Feature selection was performed and used to generate HCA dendrograms (Figure 7.2), depicting the spectral similarity between MSMT clusters. Isolates making up the MSMT clusters CL2 and CL3-CL4 being combined into one cluster by ATR-FTIR spectroscopy. While CL1 derived from MSMT data contained 14 isolates, HCA of the ATR-FTIR data placed these 14 isolates in multiple smaller clusters, highlighting additional phenotypic differences among the isolates.

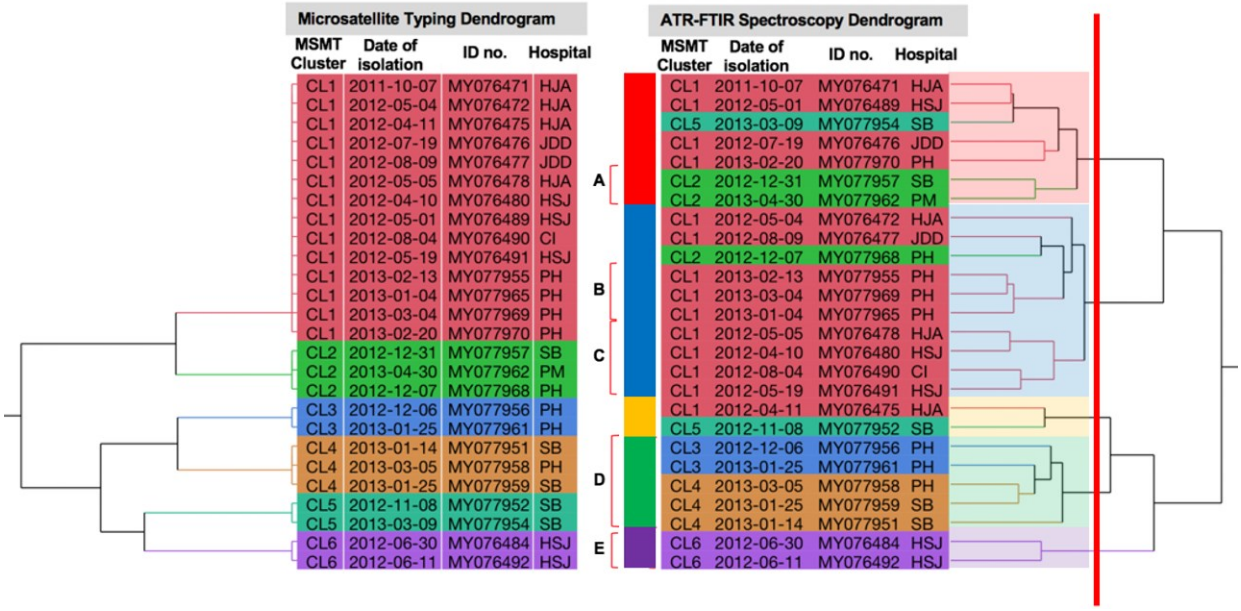


Figure 7.2. Comparison between Chilean *C. parapsilosis* suspected outbreak isolates using microsatellite (MSMT) markers and ATR-FTIR spectroscopy spectra by hierarchical cluster analysis using the spectral. (A) shows the MSMT-derived cluster is divided with CL2 cluster, (B) Isolates clustering with like-hospital within 2 months. (C) Isolates clustering with different hospitals within 4 months, (D) MSMT-derived clusters CL3 and CL4 clustered together and (E) Same clustering as MSMT. MSMT-derived CL5 is not observed by ATR-FTIR spectroscopy

From Figure 7.2, using a threshold of 65% similarity for ATR-FTIR analysis, section (A) shows the MSMT-derived cluster is divided with CL2 cluster, (B) Isolates clustering with like-hospital within 2 months. (C) Isolates clustering with different hospitals within 4 months, (D) MSMT-derived clusters CL3 and CL4 clustered together and (E) Same clustering as MSMT. MSMT-derived CL5 is not observed by ATR-FTIR spectroscopy.

Utilizing only the selected 10 isolates, similar clustering is observed and described in section 7.4.7.

7.4.6.3. ^{1}H HR-MAS NMR spectroscopy

Non-destructive, whole-organism fingerprinting ^{1}H HR-MAS NMR method was employed to provide spectra of 10 potential *C. parapsilosis* outbreak strains that were typed by WGS and ATR-FTIR spectroscopy.

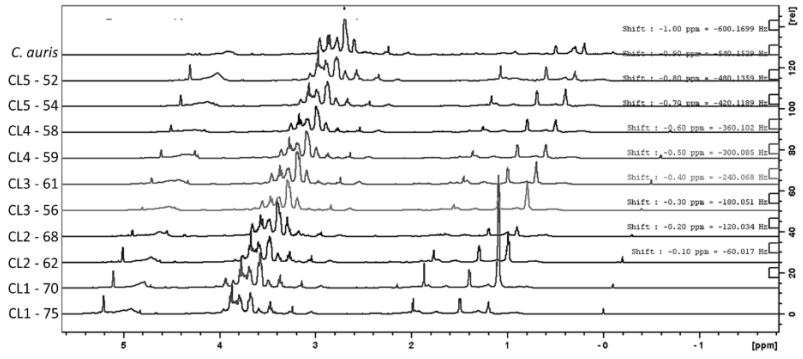


Figure 7.3. ^{1}H HR-MAS NMR spectra of outbreak *C. parapsilosis* isolates labeled by MSMT cluster and last 2 digits of their isolate identification number. Spectrum of *Candida auris* is present for visualization purposes of difference in species by HR-MS NMR spectroscopy.

Through HCA analysis, MSMT clusters CL3, CL4 and CL5 were also observed by analysis of the NMR spectra using region between 0 and 3.0 ppm and 4.2 to 5.4 ppm. These spectral regions are associated with multiple metabolites such as lipids, and amino acid (e.g., arginine, serine, threonine), ethanol, glucose and more (Figure 7.4). ^{1}H HR-MAS NMR is an extremely powerful technique that can help elucidate the nature of the chemical compounds responsible for the discrimination of the *C. parapsilosis* isolates by FTIR spectroscopy. This approach should be vigorously pursued in the future and is beyond the scope of the current thesis work.

7.4.7. Comparison of all strain-typing methods

By comparing MSMT and ATR-FTIR clusters to the Newick phylogeny tree generated from the WGS data, WGS clusters are agreeing with both MSMT and ATR-FTIR clusters (Figure 7.4). Interestingly, CL1 and CL2 in the WGS Newick tree, corresponds to those generated by the ATR-FTIR-based typing method (grouped together) while MSMT clustered into two different genotypes. Furthermore, by ATR-FTIR spectroscopy, MSMT CL5 is clustering with CL1 and CL2 while the two (CL5 and CL1-CL2) MSMT genotypic profiles share no common alleles. By WGS, the MSMT clusters of CL1 and CL2 are in the same clusters and the two clusters and closely related to CL5 (agreeing with ATR-FTIR spectroscopy).

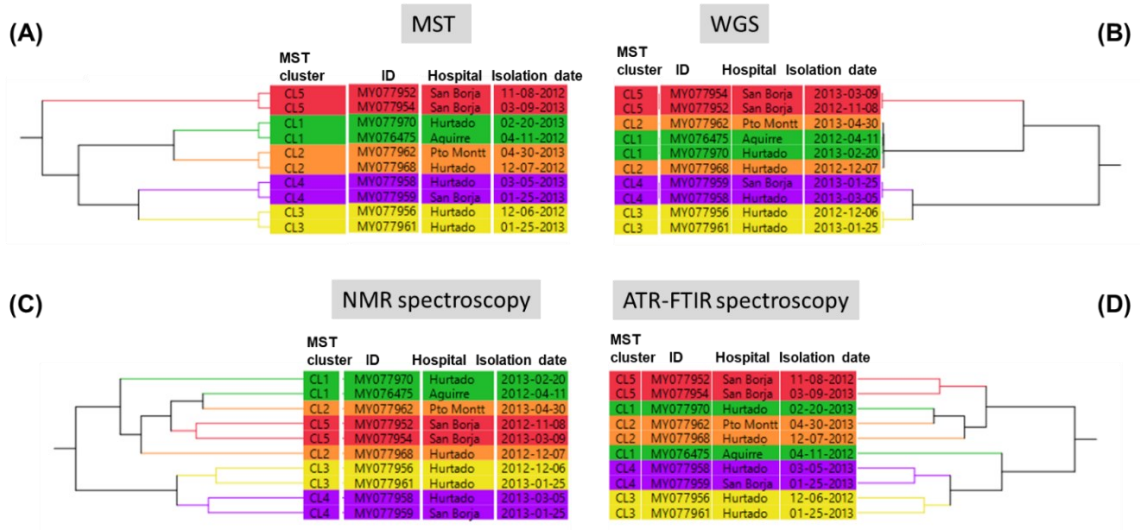


Figure 7.4. Summary of strain typing techniques; (A) microsatellite strain typing (MST), (B) whole genome sequencing (WGS), (C) high resolution magic angle spinning nuclear magnetic resonance (NMR) spectroscopy and (D) attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy. Results are displayed using a dendrogram generated from hierarchal cluster analysis of microsatellite markers, SNPs difference, ^{1}H shifts and selected wavenumbers from the ATR-FTIR spectra.

Results so far are promising for ATR-FTIR spectroscopy for strain-typing *C. parapsilosis* (a highly clonal microorganism) and WGS have so far validated close clustering of certain clusters. Moreover, increasing the microsatellite markers to more than the 5 used in the study may increase the differences between isolates within the current MSMT clusters. For example, the 2 isolates of CL1 currently have the halotype AB[AC]E[AC] using the 5

microsatellite markers, however, increasing to more markers may give rise to different halotypes from the 2 isolates. If this were true, then clustering would be in agreeance with the 2 spectroscopic techniques, suggesting the difference in phenotype of the isolates, however, by WGS, they are certainly related. Also, it should be noted that isolate MY077970 (CL1) and MY077962 (CL2) in Figure 7.4 are closely clustering together and have similar AST profiles though by ^{1}H HR-MAS NMR spectroscopy, the same clustering was not observed. Perhaps ^{13}C HR-MAS NMR may provide insight than ^{1}H for the CL1 and CL2 isolates. For all 4 strain typing methods investigated in the current chapter, MSMT clusters CL3, CL4 and CL5 were the same while WGS grouped CL1 and CL2 together and by both whole organism fingerprinting techniques, CL1 and CL2 where mixed.

Table 7.5. Summary of strain typing techniques; microsatellite strain typing, whole genome sequencing (WGS), high resolution magic angle spinning nuclear magnetic resonance (NMR) spectroscopy and attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy results of clustering

MSMT halotype and cluster		Other strain typing techniques			
Halotype	MSMT	WGS	ATR-FTIR cluster		NMR cluster
AB[AC]E[AC]	CL1	W3	A1	A2	Mix
AB[AC]EC	CL2	W3	A1		Mix
ABA[AB]A	CL3	W1	A3		N2
ABAB[BC]	CL4	W2	A4		N3
BA[AB]C[DE]	CL5	W4	A5		N4
DAAFC	CL6	-	A6		-

C. parapsilosis grown on SAB has multiple morphological forms compared to other commonly isolated yeasts such as *C. albicans, C. dubliniensis and C. tropicalis*. Only *C. albicans* compared to the other *Candida* species forms true hyphal forms (49). Unlike *C. albicans*, not all strains of *C. parapsilosis* have the capability of forming biofilms (extracellular polymeric substances) as they do not form true hyphal. Also compared to other *Candida* spp., *C. parapsilosis* produces minimal extracellular matrices (high levels of carbohydrates and low levels of protein) (7). As such, it may be interesting to examine biofilms of *C. parapsilosis* and how it related to the ATR-FTIR spectra. Glucose concentrations in the cells are directly related to biofilm formation and the increase in glucose metabolism, increases the upregulation of the pathways associated with glucose

metabolism which in turn may increase the biofilm formation and antifungal resistance (50, 51).

7.5. CONCLUSION

While ATR-FTIR spectroscopy did not provide identical results to MSMT, it was the first evaluation for whole organism fingerprinting strain typing of *C. parapsilosis*. The results of the study suggest discrimination of potential outbreak strains may be possible through spectrotyping and how ATR-FTIR spectra of microorganisms may be based on both phenotypic and genotypic information. The current study is considered as a preliminary work to explore relationship between different typing methods. More work is required to conclude on the feasibility of the method for strain typing for tracking potential outbreaks, however, if validated, the technology and technique will improve turnaround time compared with genotyping and provide resource limited laboratories with an inexpensive tool for strain typing.

7.6. REFERENCES

1. Thomaz DY, de Almeida Jr JN, Lima GME, Nunes MdO, Camargo CH, Grenfell RdC, Benard G, Del Negro G. 2018. An azole-resistant Candida parapsilosis outbreak: clonal persistence in the intensive care unit of a Brazilian teaching hospital. Frontiers in microbiology 9:2997.

2. Pinhati HMS, Casulari LA, Souza ACR, Siqueira RA, Damasceno CMG, Colombo AL. 2016. Outbreak of candidemia caused by fluconazole resistant Candida parapsilosis strains in an intensive care unit. BMC infectious diseases 16:433.

3. Mesini A, Mikulska M, Giacobbe DR, Del Puente F, Gandolfo N, Codda G, Orsi A, Tassinari F, Beltramini S, Marchese A. 2020. Changing epidemiology of candidaemia: Increase in fluconazole-resistant Candida parapsilosis. Mycoses.

4. Tóth R, Nosek J, Mora-Montes HM, Gabaldon T, Bliss JM, Nosanchuk JD, Turner SA, Butler G, Vágvölgyi C, Gácser A. 2019. Candida parapsilosis: from genes to the bedside. Clinical microbiology reviews 32:e00111-18.

5. Trofa D, Gácser A, Nosanchuk JD. 2008. Candida parapsilosis, an emerging fungal pathogen. Clinical microbiology reviews 21:606-625.

6. Reiss E, Shadomy HJ, Lyon GM. 2011. Fundamental medical mycology. John Wiley & Sons.

7. Cavalheiro M, Teixeira MC. 2018. Candida biofilms: threats, challenges, and promising strategies. Frontiers in medicine 5:28.

8. de Paula Menezes R, de Oliveira Melo SG, Bessa MAS, Silva FF, Alves PGV, Araújo LB, Penatti MPA, Abdallah VOS, de Brito Röder DvD, dos Santos Pedroso R. 2020. Candidemia by Candida parapsilosis in a neonatal intensive care unit: human and environmental reservoirs, virulence factors, and antifungal susceptibility. Brazilian Journal of Microbiology:1-10.

9. Qi L, Fan W, Xia X, Yao L, Liu L, Zhao H, Kong X, Liu J. 2018. Nosocomial outbreak of Candida parapsilosis sensu stricto fungaemia in a neonatal intensive care unit in China. Journal of Hospital Infection 100:e246-e252.

10. Hernández-Castro R, Arroyo-Escalante S, Carrillo-Casas EM, Moncada-Barrón D, Álvarez-Verona E, Hernández-Delgado L, Torres-Narváez P, Lavalle-Villalobos A. 2010. Outbreak of Candida parapsilosis in a neonatal intensive care unit: a health care workers source. European journal of pediatrics 169:783-787.

11. Madhavan P, Chong P, Farida J, Fauziah O, Cheah Y, Arunkumar Karunanidhi A, Ng K. 2019. Antimicrobial susceptibilities and random amplified polymorphic DNA-PCR fingerprint characterization of Candida glabrata, Candida parapsilosis and Candida rugosa from two major hospitals in Kuala Lumpur, Malaysia. Tropical Biomedicine 36:183-193.

12. Desnos-Ollivier M, Bórmida V, Poirier P, Nourrisson C, Pan D, Bretagne S, Puime A, Dromer F, Network UIFI, Group FMS. 2018. Population structure of Candida parapsilosis: No genetic difference between French and Uruguayan isolates using microsatellite length polymorphism. Mycopathologia 183:381-390.

13. Badali H, Rezaie S, Meis JF, Afshari SAK, Modiri M, Hagen F, Moazeni M, Mohammadi R, Khodavaisy S. 2017. Microsatellite genotyping of clinical Candida parapsilosis isolates. Current medical mycology 3:15.

14. Arvanitis M, Anagnostou T, Fuchs BB, Caliendo AM, Mylonakis E. 2014. Molecular and nonmolecular diagnostic methods for invasive fungal infections. Clinical microbiology reviews 27:490-526.

15. Barbedo LS, Figueiredo-Carvalho MHG, Muniz MdM, Zancopé-Oliveira RM. 2017. Comparison of four molecular approaches to identify Candida parapsilosis complex species. Memórias do Instituto Oswaldo Cruz 112:214-219.

16. Selmecki A, Forche A, Berman J. 2010. Genomic plasticity of the human fungal pathogen Candida albicans. Eukaryotic cell 9:991-1008.

17. Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, Colombo AL, Calvo B, Cuomo CA, Desjardins CA. 2016. Simultaneous emergence of multidrug-resistant Candida auris on 3 continents confirmed by whole-genome sequencing and epidemiological analyses. Clinical Infectious Diseases 64:134-140.

18. Zoll J, Snelders E, Verweij PE, Melchers WJ. 2016. Next-generation sequencing in the mycology lab. Current fungal infection reports 10:37-42.

19. Lam LM, Dufresne PJ, Longtin J, Sedman J, Ismail AA. 2019. Reagent-Free Identification of Clinical Yeasts by Use of Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy. Journal of clinical microbiology 57:e01739-18.

20. Costa FS, Silva PP, Morais CL, Arantes TD, Milan EP, Theodoro RC, Lima KM. 2016. Attenuated total reflection Fourier transform-infrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii. Analytical Methods 8:7107-7115.

21. Silva S, Tobaldini-Valerio F, Costa-de-Oliveira S, Henriques M, Azeredo J, Ferreira EC, Lopes JA, Sousa C. 2016. Discrimination of clinically relevant Candida species by Fourier-transform infrared spectroscopy with attenuated total reflectance (FTIR-ATR). RSC Advances 6:92065-92072.

22. Zarnowiec P, Lechowicz L, Czerwonka G, Kaca W. 2015. Fourier transform infrared spectroscopy (FTIR) as a tool for the identification and differentiation of pathogenic bacteria. Current medicinal chemistry 22:1710-1718.

23. Naumann A. 2015. Fourier transform infrared (FTIR) microscopy and imaging of fungi, p 61-88, Advanced Microscopy in Mycology. Springer.

24. Lecellier A, Gaydou V, Mounier J, Hermet A, Castrec L, Barbier G, Ablain W, Manfait M, Toubas D, Sockalingum G. 2015. Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds. Food microbiology 45:126-134.

25. Wenning M, Breitenwieser F, Konrad R, Huber I, Busch U, Scherer S. 2014. Identification and differentiation of food-related bacteria: a comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry. Journal of microbiological methods 103:44-52.

26. Schmitt J, Flemming H-C. 1998. FTIR-spectroscopy in microbial and material analysis. 41:1-11.

27. Vogt S, Löffler K, Dinkelacker AG, Bader B, Autenrieth IB, Peter S, Liese J. 2019. Fourier-transform infrared (FTIR) spectroscopy for typing of clinical Enterobacter cloacae complex isolates. Frontiers in microbiology 10.

28. Novais Â, Freitas AR, Rodrigues C, Peixe L. 2019. Fourier transform infrared spectroscopy: unlocking fundamentals and prospects for bacterial strain typing. European Journal of Clinical Microbiology & Infectious Diseases 38:427-448.

29. Johler S, Stephan R, Althaus D, Ehling-Schulz M, Grunert T. 2016. High-resolution subtyping of Staphylococcus aureus strains by means of Fourier-transform infrared spectroscopy. Systematic and applied microbiology 39:189-194.

30. Davis R, Paoli G, Mauer L. 2012. Evaluation of Fourier transform infrared (FT-IR) spectroscopy and chemometrics as a rapid approach for sub-typing Escherichia coli O157: H7 isolates. Food microbiology 31:181-190.

31. Toubas D, Essendoubi M, Adt I, Pinon J-M, Manfait M, Sockalingum GD. 2007. FTIR spectroscopy in medical mycology: applications to the differentiation and typing of Candida. Analytical and bioanalytical chemistry 387:1729.

32. Essendoubi M, Toubas D, Lepouse C, Leon A, Bourgeade F, Pinon J-M, Manfait M, Sockalingum GD. 2007. Epidemiological investigation and typing of Candida glabrata clinical isolates by FTIR spectroscopy. Journal of microbiological methods 71:325-331.

33. Sandt C, Sockalingum G, Aubert D, Lepan H, Lepouse C, Jaussaud M, Leon A, Pinon J, Manfait M, Toubas D. 2003. Use of Fourier-transform infrared spectroscopy for typing of Candida albicans strains isolated in intensive care units. Journal of clinical microbiology 41:954-959.

34. Taha M, Hassan M, Essa S, Tartor Y. 2013. Use of Fourier transform infrared spectroscopy (FTIR) spectroscopy for rapid and accurate identification of yeasts isolated from human and animals. International journal of veterinary science and medicine 1:15-20.

35. Sandt CL, Sockalingum GD, Toubas D, Aubert D, Lepan H, Lepouse C, Jaussaud M, Leon A, Pinon J-M, Manfait M. Comparing FTIR and RAPD techniques in the

typing of c. albicans in a clinical set-up, p 1-11. *In* (ed), International Society for Optics and Photonics,

36. Corte L, di Cagno R, Groenewald M, Roscini L, Colabella C, Gobbetti M, Cardinali G. 2015. Phenotypic and molecular diversity of Meyerozyma guilliermondii strains isolated from food and other environmental niches, hints for an incipient speciation. Food microbiology 48:206-215.

37. Hanoulle X, Wieruszeski J-M, Rousselot-Pailley P, Landrieu I, Baulard AR, Lippens G. 2005. Monitoring of the ethionamide pro-drug activation in mycobacteria by 1H high resolution magic angle spinning NMR. Biochemical and biophysical research communications 331:452-458.

38. Li W. 2006. Multidimensional HRMAS NMR: a platform for in vivo studies using intact bacterial cells. Analyst 131:777-781.

39. Gudlavalleti SK, Szymanski CM, Jarrell HC, Stephens DS. 2006. In vivo determination of Neisseria meningitidis serogroup A capsular polysaccharide by whole cell high-resolution magic angle spinning NMR spectroscopy. Carbohydrate research 341:557-562.

40. Righi V, Constantinou C, Kesarwani M, Rahme LG, Tzika AA. 2018. Effects of a small, volatile bacterial molecule on Pseudomonas aeruginosa bacteria using whole cell high-resolution magic angle spinning nuclear magnetic resonance spectroscopy and genomics. International journal of molecular medicine 42:2129-2136.

41. Righi V, Constantinou C, Kesarwani M, Rahme LG, Tzika AA. 2013. Live-cell high resolution magic angle spinning magnetic resonance spectroscopy for in vivo analysis of Pseudomonas aeruginosa metabolomics. Biomedical reports 1:707-712.

42. Himmelreich U, Somorjai RL, Dolenko B, Daniel H-M, Sorrell TC. 2005. A rapid screening test to distinguish between Candida albicans and Candida dubliniensis using NMR spectroscopy. FEMS microbiology letters 251:327-332.

43. Himmelreich U, Somorjai RL, Dolenko B, Lee OC, Daniel H-M, Murray R, Mountford CE, Sorrell TC. 2003. Rapid identification of Candida species by using nuclear magnetic resonance spectroscopy and a statistical classification strategy. Appl Environ Microbiol 69:4566-4574.

44. Bourne R, Himmelreich U, Sharma A, Mountford C, Sorrell T. 2001. Identification of Enterococcus, Streptococcus, andStaphylococcus by Multivariate Analysis of Proton Magnetic Resonance Spectroscopic Data from Plate Cultures. Journal of clinical microbiology 39:2916-2923.

45. Tintelnot K, Haase G, Seibold M, Bergmann F, Staemmler M, Franz T, Naumann D. 2000. Evaluation of phenotypic markers for selection and identification of Candida dubliniensis. Journal of clinical microbiology 38:1599-1608.

46. Lasker BA, Butler G, Lott TJ. 2006. Molecular genotyping of Candida parapsilosis group I clinical isolates by analysis of polymorphic microsatellite markers. Journal of clinical microbiology 44:750-759.

47. Reiss E, Lasker BA, Lott TJ, Bendel CM, Kaufman DA, Hazen KC, Wade KC, McGowan KL, Lockhart SR. 2012. Genotyping of Candida parapsilosis from three neonatal intensive care units (NICUs) using a panel of five multilocus microsatellite markers: broad genetic diversity and a cluster of related strains in one NICU. Infection, Genetics and Evolution 12:1654-1660.

48. Tavanti A, Davidson AD, Gow NA, Maiden MC, Odds FC. 2005. Candida orthopsilosis and Candida metapsilosis spp. nov. to replace Candida parapsilosis groups II and III. Journal of clinical microbiology 43:284-292.

49. Holland LM, Schröder MS, Turner SA, Taff H, Andes D, Grózer Z, Gácser A, Ames L, Haynes K, Higgins DG. 2014. Comparative phenotypic analysis of the major fungal pathogens Candida parapsilosis and Candida albicans. PLoS pathogens 10:e1004365.

50. Pérez-García LA, Csonka K, Flores-Carreón A, Mellado-Mojica E, Németh T, López-Ramírez LA, Toth R, López MG, Vizler C, Marton A. 2016. Role of protein

glycosylation in Candida parapsilosis cell wall integrity and host interaction. Frontiers in microbiology 7:306.

51. Pereira L, Silva S, Ribeiro B, Henriques M, Azeredo J. 2015. Influence of glucose concentration on the structure and quantity of biofilms formed by Candida parapsilosis. FEMS Yeast Research 15:fov043.

Connecting statement 6

The following section discusses the major findings of the previous chapters and implications for the implementation of the ATR-FTIR spectrotyping technique for microbial analysis.

CHAPTER 8. GENERAL DISCUSSION

The research work presented in this thesis is aimed at the development, standardization, and evaluation of a novel spectrotyping approach using attenuated total reflectance Fourier transform (ATR-FTIR) spectroscopy as a microbial identification and strain-typing technique. In this work, ATR-FTIR spectrotyping of over 200 species (from over 70 genera) isolated from clinical specimens and food sources was conducted. The results of this work successfully demonstrated the potential applicability of ATR-FTIR spectrotyping for microbial identification at the Gram-stain type, genus, species, strain, and serotype levels.

Beyond providing evidence supporting these potential applications, the research work presented in this thesis has demonstrated that the ATR-FTIR spectrotyping technique can be readily integrated in current clinical and food microbiology lab workflows. Conventional biochemical techniques for species identification may not work for all microorganisms and may require that genetic sequencing be performed to confirm species identification. The ATR-FTIR spectrotyping technique may potentially be used to speciate various organisms once conventional biochemical techniques have been used to determine the genera. This type of application will greatly improve time-to-result and requires no reagents. Furthermore, the methodology is extremely simple and can easily be implemented in laboratories of all sizes. The ATR-FTIR spectrotyping technique may also have a potential role as an adjunct to matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS), which is well known for its inability to differentiate between phylogenetically similar microorganisms (such as *Shigella* spp. and *Escherichia coli*) using the mass range relating to ribosomal proteins. The ATR-FTIR spectrotyping technique may also potentially serve as a rapid screening tool for outbreak detection.

For the advancement of ATR-FTIR spectroscopy in clinical microbiology, the current thesis provides standardized operating procedures for the growth of the microorganisms. Slight formulation differences of the same type of agar culture media may affect spectral reproducibility and result in low confidence or misidentification of the microorganism. Standard reference strains may be used as a quality control check to ensure the spectral

profile of the microorganism is similar to the spectrum of the identical reference strain in the spectral database. To provide a more robust methodology, it may be possible to construct spectral databases to represent microorganisms grown on various agar culture media and/or grown under various culturing conditions. Another factor to consider for implementation of the methodology would be to validate different makes and models of spectrometers. Not all spectrometers are built the same and the spectra may vary in wavenumber alignment and signal-to-noise ratio. A spectral database constructed using a single spectrometer may not provide accurate results from spectra collected from a spectrometer of a different model. Similar to the culture media effect, it may be possible using machine learning algorithms to train the prediction model using spectra from multiple spectrometers.

The novel spectrotyping method employed a myriad of multivariate analysis techniques such as principal component analysis, hierarchical cluster analysis and support vector machine in combination with the use of targeted spectral regions. For microbial identification, multiple multivariate models may be combined to create a single-step microbial identification method at all the taxonomic levels listed above. Numerous studies published in the literature have explored various multivariate statistical analysis tools but have yet to directly compare various software and data analysis methods using a large set of infrared spectral data of diverse bacteria and yeasts. Evaluation of the performances of different algorithms will go towards the development of an expert artificial intelligence-based microbial identification system by combining the results of each algorithm and by consensus to achieve an accurate microbial identification platform. Further work can be pursued for the evaluation of different multivariate analysis algorithms and combine selected algorithms to improve predictive accuracy of the ATR-FTIR spectroscopy-based method for microbial identification and strain typing.

Phenotypic properties of microorganisms are dependent on multiple environmental factors such as the atmospheric conditions, carbon and/or nitrogen source, presence of antimicrobials, salts, and metal ions in the surrounding medium, and pH. With changes in the surrounding environment, changes in the microorganism’s metabolism as well as changes in cell wall composition may take place in order for the microorganism to adapt

to the new environment (reiterating the importance of a standard agar culture media). The adaptation to the new environment is the result of changes in gene expression such as turning on or off genes to produce certain proteins and/or make changes to the cellular structure (especially the cellular membrane). That is to say, the genetic code stays the same while the phenotype changes. Spectrotyping is a unique spectral classification technique solely based on the biochemical composition of a microorganism. As such, the evolution of microorganisms (as well as changes in their classification based on genotypic data) may impact the performance of the spectrotyping technique for microbial identification. Accordingly, spectral databases will also have to evolve over time. Increasing the spectral representation of microorganisms from various geographical locations and current strains will have to be done. With the need for continual spectral database updates, it is uncertain how many isolates are needed for adequate spectral representation of one class (i.e., genus, species, serotype). Based on the results of the present study, some species may need spectral representation of 5 isolates to provide accurate results while other species require more than 20. Adequate spectral representation may be species dependent, and phenotypic and genotypic attributes may aid in determining the extent of spectral diversity within a given species. For example, if a species is known to rapidly evolve and displays varying phenotypes when grown on different agar culture media, many more than 20 isolates may be required for adequate spectral representation of this particular species.

For epidemiology, pulsed-field gel electrophoresis is the current gold standard for genotyping and detecting outbreaks, but it can be time-consuming and costly, requiring intensive training and experience, and sometimes may not be fully transferable between laboratories. ATR-FTIR spectrotyping can potentially provide resource-limited laboratories with the ability to strain-type microorganisms for surveillance purposes. As such, correlation between genotypic and phenotypic attributes between microorganisms will need to be further investigated in relation to the observed differences in the ATR-FTIR spectra of closely related strains. In this regard, it may be noted that there are many factors to consider when trying to relate whole genome sequencing data with phenotypic data as some loci on the genome are non-coding with no significant functional significance (need to validate regions), multiple genes working together to produce a

phenotype, one single nucleotide polymorphism (SNP) causing extreme or no phenotypic changes, combination of SNPs to alter phenotypic expression, and SNP altering the expression of another gene.

Spectrotyping can shed light on the nature of the biochemical differences between species or strains; however, additional analytical techniques that provide superior chemical specificity will be required to identify specific biomarkers. Elucidation of the chemical structure of biomarkers responsible for the spectral differences on which discrimination between particular species or strains is based, through the application of MALDI-TOF MS, liquid chromatography-tandem mass spectrometry and nuclear magnetic resonance spectroscopy, may play an important role in increasing confidence in ATR-FTIR spectrotyping as a rapid technique for microbial identification and strain typing.

CHAPTER 9. CONCLUSION AND SUMMARY

Fourier transform infrared (FTIR) spectroscopy is a well-established analytical technology that has been used in various fields of study such as chemistry, forensics, and medicine. In microbiology, the infrared (IR) spectrum of microorganisms acquired directly from initial culture is representative of their biochemical composition and is referred to as a whole-organism fingerprint. FTIR spectroscopy coupled with the attenuated total reflectance (ATR) mode of spectral acquisition greatly simplifies spectral acquisition, thus offering an easy-to-use, cost-effective, and sensitive analytical technology.

The present work addresses the potential applicability of ATR-FTIR spectroscopy in routine microbiology for the identification and typing of bacteria and yeasts. At the outset of this work, the elaboration of a standardized operating protocol was necessary to ensure that the methodologies developed would be directly transferable to external sites for validation and potentially routine implementation. This part of the work entailed assessment of instrument-to-instrument (make and models) spectral variability, microbial sample preparation (effects of culture media and incubation conditions), sample deposition approach that ensures controlled and consistent sample humidity, and spectral acquisition parameters impacting signal-to-noise ratio. In addition, spectral pre-processing steps compensating for variability in ambient atmospheric moisture were established.

Following development of a standardized operating protocol, ATR-FTIR-based “spectrotyping” was undertaken. Spectrotyping is a novel approach centered on the process of determining the differences between spectra (of microorganisms) based on the absence or presence and relative intensities of IR absorption bands and was assessed in the present work for differentiation of microorganisms at the species level. For this purpose, food, and clinical isolates of bacteria (*n*=2619) and yeasts (*n*=391) were obtained from the frozen collections of 9 microbiology laboratories. These isolates had all been identified by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) and conventional biochemical identification techniques and belonged to 20 genera of bacteria and 12 genera of yeasts, accounting for a total of 92 species of Gram-negative (*n*=14) and Gram-positive bacteria (*n*=13) and yeasts

(*n*=65). In accordance with the standardized operating protocol, ATR-FTIR spectra of the bacterial isolates were acquired in triplicate by directly transferring colonies from the culture plate onto the ATR sampling surface. Averaged triplicate spectra of the 2619 isolates were divided into training and test sets where isolates in the test set were selected based on stratified random sampling of ~30% of the total isolates in each group (species). Applying a multitude of multivariate spectral analysis methods to the ATR-FTIR spectra in the training set (e.g., k nearest neighbor in conjunction with hierarchical cluster analysis, principal component analysis) and support vector machine (SVM), a sequential pairwise multitier prediction model was constructed. Predictions of Gram-stain type, genus, and species were then obtained from this model for the spectra in the test set. The prediction model yielded 99.3% and 99.4% correct Gram-stain type identification for Gram-negative (*n*=272) and Gram-positive bacteria (*n*=312), respectively. At the genus level, 100% correct identification was achieved for clinically relevant microorganisms such as *Shigella* spp., *Enterococcus* spp., *Listeria* spp., and *Staphylococcus* spp.; overall, 94.3% correct genus-level identification was achieved for all genera represented in the training set (at least 14 strains per species within each genus). Moreover, at the species level, 92.6% and 99.1% correct identification was obtained for Gram-negative and Gram-positive bacteria, respectively. Importantly, within the Gram-negative group, 100% correct species identification was achieved for *E. coli* (*n*=31), *Pseudomonas aeruginosa* (*n*=47), *Salmonella enterica* (*n*=42), *Shigella flexneri* (*n*=6) and *Shigella sonnei* (*n*=12). All isolates belonging to major Gram-positive species [*Enterococcus faecium* (*n*=22), *Enterococcus faecalis* (*n*=26), *Listeria monocytogenes* (*n*=13), *Staphylococcus aureus* (*n*=110) and species within coagulase-negative staphylococci (*n*=43)] were correctly identified.

An external validation study of the ATR-FTIR spectroscopy-based bacteria identification method employing the prediction model developed in this work was conducted at one of the participating hospitals. Over a 3-month period, a total of 391 bacterial isolates were collected directly from culture plates in the routine microbiology laboratory (representative of typical species distribution on a daily basis) and were identified by the ATR-FTIR spectroscopy-based method in parallel with automated conventional biochemical techniques and/or MALDI-TOF MS. Among the 31 species identified by the routine

methods, 19 species were not represented in the training set employed to develop the prediction model but accounted for only 28 isolates. Among the other 363 isolates, belonging to species represented in the training set, 98.7%, 91.3% and 98.4% were correctly identified at the Gram-stain type, genus, and species level, respectively. Furthermore, the 28 isolates belonging to the non-represented species all gave a result of “inconclusive” rather than being misidentified. The development of a sequential pairwise multitier prediction model based on spectrotyping of yeasts was undertaken with a training set initially consisting of 199 frozen isolates (previously identified by reference methods) encompassing 5 genera and 14 species. The performance of the prediction model developed with this training set was evaluated in a 4-month external validation study in which fresh isolates (*n*=318) from routine clinical specimens were collected directly from culture plates and identified by the ATR-FTIR spectroscopy-based method in parallel with MALDI-TOF MS and/or target PCR gene sequencing. Correct species identification of 95.6% of the isolates was obtained; the misidentification rate was 0.9% (*n*=3) while 11 isolates, gave a result of “inconclusive”. Following optimization of the prediction model with the addition of 62 isolates to the training set, resulting in representation of 12 genera and 65 species, re-analysis of the spectra of the 318 routine isolates resulted in 99.7% correct species identification with a reduced misidentification rate of 0.3% (*n*=1). Subsequently, the ATR-FTIR spectroscopy-based method for yeast identification employing the optimized prediction model was evaluated in a multicenter study (unsupervised) encompassing 6 clinical microbiology laboratories using 3 different spectrometers of the same model, resulting in 98.3% correct species identification with no misidentification of 534 collected isolates.

Strain typing of yeasts based on infrared spectral profiles was also examined in the present work in relation to a retrospective investigation of a potential *Candida parapsilosis* outbreak in Chile. This study entailed typing of 41 isolates from the suspected outbreak and enabled a direct comparison between the results obtained by the ATR-FTIR spectroscopy-based technique and those obtained by whole-genome sequencing and microsatellite molecular typing.

The research work presented in this thesis established the potential of employing ATR-FTIR spectroscopy as a routine technique for microbial identification. The validation studies of the ATR-FTIR spectroscopy-based methods for bacteria and yeasts based on the prediction models developed in this research demonstrated comparable results to well-accepted MALDI-TOF MS, paving the way for future accreditation. The implementation of this reagent-free and rapid technique could be of particular benefit to smaller-scale microbiology laboratories as a cost-effective alternative to MALDI-TOF MS.

APPENDIX

Table A.1. Table of classification levels of Gram-negative bacteria commonly isolated from clinical and food microbiology for the evaluation of ATR-FTIR-based identification and discrimination method

Microorganism	Bacteria			
Gram-stain	Gram-negative			
Genus	*Salmonella*	*Escherichia*	*Klebsiella*	*Shigella*
Species	*S. enterica*	*E. coli*	*K. pneumoniae* *K. oxytoca*	*S. sonnei* *S. flexneri*
Serotype	Heidelberg Typhimurium Thompson Hadar Newport Enteritidis	O157:H7 non-O157:H7		
Gram-stain	Gram-negative			
Genus	*Citrobacter*	*Enterobacter*	*Achromobacter*	*Acinetobacter*
Species	*C. braakii* *C. freundii* *C. koseri*	*E. cloacae* *E. kobei*	*A. denitrificans* *A. xylosoxidans*	*A. baumannii*
Non-fermenting Gram-negative bacilli				
Genus	*Burkholderia*		*Stenotrophomonas*	*Pseudomonas*
Group	Non-BCC	BCC		
Species	*B. gladioli*	*B. ambifaria* *B. anthina* *B. cenocepacia* *B. cepacia* *B. multivorans* *B. vietnamensis*	*S. maltophilia*	*P. aeruginosa*

Table A.2. Table of classification levels of Gram-positive bacteria commonly isolated from clinical and food microbiology for the evaluation of ATR-FTIR-based identification and discrimination method

Microorganism	Bacteria					
Gram-stain	Gram-positive					
Genus	*Staphylococcus*			*Enterococcus*		*Listeria*
Group	Non-CoNS		CoNS			
Species	*S. aureus*		*S. capitis* *S. epidermidis* *S. haemolyticus* *S. hominis* *S. lugdunensis* *S. warneri*	*E. faecium* *E. faecalis* *E. gallinarum*		*L. monocytogenes* *L. innocua* *L. ivanovii* *L. marhii* *L. seeligeri* *L. welshimeri*
Antimicrobial resistance	MRSA	MSSA		VRE	VSE	

Genus	*Streptococcus*				*Bacillus*
Group	Group A	Group B	Group G	*S. viridans* grp.	
Species	*S. pyogenes*	*S. agalactiae*	*S. dysgalactiae*	*S. anginosus* *S. mitis* *S. sanguinis* *S. salivarius* *S. mutans*	*B. cereus* *B. thuringiensis*

Table A.3. Table of classification levels of yeasts commonly isolated from clinical and food microbiology for the evaluation of ATR-FTIR-based identification and discrimination method

<table>
<tr><td>Kingdom</td><td colspan="3">Fungi</td></tr>
<tr><td>Microorganism</td><td colspan="3">Yeast</td></tr>
<tr><td>Genus</td><td colspan="2">Candida</td><td>Cryptococcus</td></tr>
<tr><td>Species</td><td colspan="2">C. albicans
C. auris
C. dubliniensis
C. duobushaemulonii
C. glabrata
C. guilliermondii
C. haemulonii
C. kefyr
C. krusei
C. lipolytica
C. lusitaniae
C. orthopsilosis
C. parapsilosis
C. pararugosa
C. pelliculosa
C. tropicalis
C. utilis</td><td>C. neoformans</td></tr>
<tr><td rowspan="2">Antimicrobial resistance</td><td>Fluconazole resistance</td><td>Fluconazole sensitive</td><td></td></tr>
<tr><td colspan="2">C. auris</td><td></td></tr>
<tr><td>Epidemiology strain-typing</td><td colspan="2">C. parapsilosis</td><td></td></tr>
</table>

Table A.4. Composition of various culture media agar for the evaluation of spectral variation of *Candida albicans*

Media/manufacturer/final pH	Ingredients	(g/L)	Notes:
Sabouraud Dextrose Agar (SAB)/Difco/5.6+/-0.2 @ RT	Peptic digest of Animal Tissue	5	Standard yeast nutrient agar
	Pancreatic digest of Casein	5	
	Dextrose	40	
	Agar	15	
TSA 5% Sheep blood (Blood)/Oxoid/7.3+/-0.2 @ 25C	Agar	12	General nutrient agar
	Tryptone	14	
	Peptone Neutralized	4.5	
	Yeast extract	4.5	
	Sodium chloride	5	
Candida chromogenic agar (CHROME)/CHROMagar/6.1+/-0.2 @15-30C	Agar	15	*C. albicans* = green colonies
	Peptone	10.2	*C. tropicalis* = metallic blue
	Chromogenic mix	22	*C. krusei* = pink, fuzzy
	Chloramphenicol	0.5	Other species = white to mauve
Inhibitory Mold Agar (IMA)/BD-BBL	Pancreatic Digest of Casein	3	Inhibits mold and bacterial growth to promote yeasts
	Peptic Digest of Animal Tissue	2	
	Yeast Extract	5	
	Dextrose	5	
	Starch	2	
	Dextrin	1	
	Sodium Phosphate	2	
	Magnesium Sulfate	0.8	
	Ferrous Sulfate	0.04	
	Sodium Chloride	0.04	
	Manganese Sulfate	0.16	
	Agar	15	
	Chloramphenicol	0.125	
Potato Dextrose Agar (PDA)/Difco/5.6+/-0.2 @ RT	Potato starch (from infusion)	4	General nutrient for mold agar
	Dextrose	20	
	Agar	15	

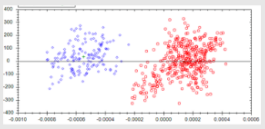
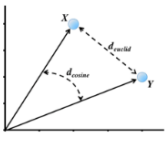
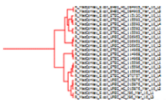
Transforms	Details	Examples
Principal component analysis (PCA)	A mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible	2D-PC-Plot X=PC2 Y=PC3 
Hierarchical cluster analysis (HCA)	Spectral distance metric between vectors x and y: Cosine $d_{cosine} = cos^{-1}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=0}^{n} x_i^2 \sum_{i=0}^{n} y_i^2}}\right)$ Linkage method: Ward (total within-cluster sum of squares as a result of joining C_1 and C_2) $d(C_1, C_2) = \frac{\lvert C_1 \rvert C_2 \rvert d(C_1, C_2)}{(\lvert C_1 \rvert + \lvert C_2 \rvert)}$	 
Greedy best-first search	A greedy search examines the combination of adjacent features and follows the path of greatest improvement. In this way it is greedy in that it only considers this path of greatest improvement. Heuristic function: $f(n) = h(n)$	

Figure A.1. DataAnalysis PCA, HCA and forward search methodology descriptions

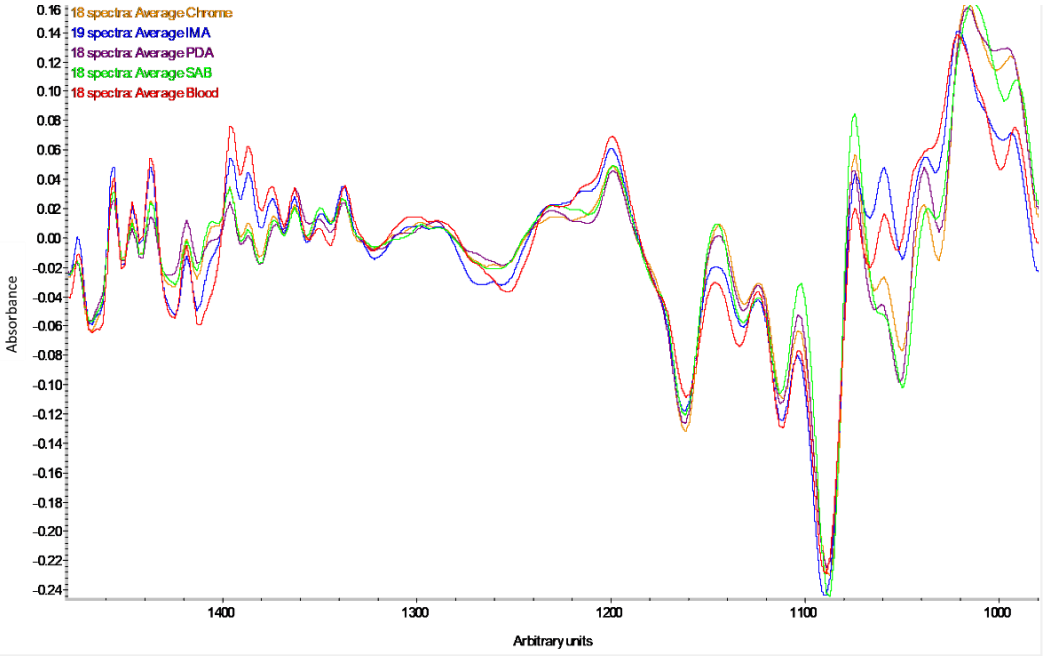


Figure A.2. Spectral variances (at 1480-980 cm^{-1} of the 1st derivative spectra) between 6 *C. albicans* isolates (averaged 3 spectra per isolate) grown on 5 different growth media

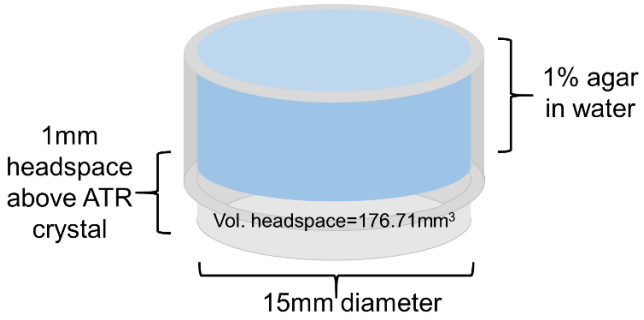


Figure A.3. Optimal 1% agar in water (0.99 A_w) cap design with a 176.71 mm^3 headspace volume between the ATR sampling surface and the agar to maintain moisture content of colonies

Table A.5. Example Standard operating procedure for ATR-FTIR spectroscopy spectral collection

Clinical Microbiology Laboratory			
Standard operating procedure for bacteria and yeast identification by the attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy-based method			
Document No.: ###-###-###			Version: YYYY-MM-DD
Manual:			Status: DRAFT
	Name(s)	Signature(s)	Date (YYYY-MM-DD)
Written by			
Revised by			
Approved by			

Document revision history			
Version	Effective date	Approved by	Description of revision

1. PRINCIPLE

Vibrational spectroscopic techniques provide complementary information on the chemical composition of a biological sample without sample processing. Traditionally associated with chemical laboratories, infrared spectroscopy is a well-established and relatively simple technique for the study and analysis of virtually any type of sample or material. With the development of Fourier transform infrared (FTIR) technology, infrared spectroscopic examination of intact cells, biological tissues, and biofluids (blood, serum, urine, and saliva) with minimal sample preparation became feasible, opening the door to biomedical applications of this spectroscopic technique.

2. PURPOSE

This protocol describes the use of an FTIR Spectrometer (Summit™, operating under OMNIC™ Paradigm™ Software provided by Thermo Scientific Inc.), procedures to prepare saliva samples for deposition onto the FTIR sampling surface, how to use the OMNIC™ Paradigm™ Software to acquire spectra and how to export spectra for data analysis using TQ Analyst software (Thermo Scientific Inc.). Spectrometer is CE marked and the software is 21 CFR Part 11 compliant and includes the data security suite (Thermo Scientific Inc.).

The procedures outlined in the document is strictly for the sample preparation and FTIR spectral acquisition of saliva samples. Data analysis is not covered in this manual and can be referred to in a separate SOP manual.

3. EQUIPMENT AND REAGENTS

Equipment

- ATR-FTIR spectrometer placed on a laboratory benchtop
  Require access to:
  - Containment level 2 and 3 laboratories with biosafety cabinet (BSC)

Materials

- Disposable 1 µL
- Lint-free tissue

Reagents

- 70% ethanol (general disinfectant for the ATR sampling surface)
- and/or 0.5% hypochlorite (to disinfect ATR sampling surface for spore forming microorganisms)

4. SAMPLE COLLECTION

Samples can be collected from hospitals, food and/or clinical reference laboratories are generally isolated from urine, sputum, positive blood cultures, pus, wounds, and swabs from various parts of the body while food-related microorganisms are isolated from animal feed and contaminated food products.

Samples for spectral acquisition can be acquired from samples collected in routine (and cultured onto culture agar media) or from long-term frozen storage samples (stored in 10% glycerol at -80°).

5. SAMPLE PREPARATION (culturing)

Frozen samples are cultured on media of choice and are subcultured once to ensure purity of the isolates, while samples collected from routine (freshly cultured from source) are analyzed as-is. Subculturing of the routine plates may also be required to ensure purity of the isolates; however, it is up to the discretion of the microbiologist. It is suggested to use the 4-streak quadrant method to obtain isolated colonies and visually assess the purity of the sample in question. If the sample obviously impure, additional subculture of the sample may be completed. According to Clinical & Laboratory Standards Institute Guidelines (CLSI, 2008), aerobic bacteria are cultured and incubated for 18-24 h at 35 ºC ± 2 ºC and 30 ºC ± 2 ºC for 24-48 h for yeasts

6. SAMPLE DEPOSITION

Sample are directly isolated from the culture agar plates using a 1 µL disposable loop and deposited onto the ATR sampling surface (diamond crystal). Per sample plate, 3 or more spectra are collected from different isolates.

7. SPECTRAL COLLECTION

A background and sample scans of 64 co-added scans at an 8 cm^{-1} resolution between 4000 and 650 cm^{-1} using the Happ-Genzel apodization function. The spectrum of the sample is ratioed against the background spectrum to yield a transmittance spectrum which is converted to an absorbance spectrum. Using these parameters, an ATR-FTIR spectrum is collected in about 1 minute, about 3 minutes per sample (3 replicates).

8. DISINFECTION

Clean the crystal with 70% ethanol (do not spay directly spray on the spectrometer – will cause aerosol. Moisten tissue with ethanol and clean the crystal)

Table A.6. Summary table of total number of collected isolates by gram-type, genus, species, and strain-type

Gram-type	Genus	Species	Strain	No. isolates collected
GN	*Achromobacter*	*denitrificans*		2
GN	*Achromobacter*	*species*		81
GN	*Achromobacter*	*xylosoxidans*		1
GN	*Acinetobacter*	*baumannii-complex*		10
GN	*Acinetobacter*	*lwoffii*		4
GN	*Acinetobacter*	*radioresistens*		2
GN	*Acinetobacter*	*species*		3
GN	*Acinetobacter*	*ursingii*		2
GN	*Aeromonas*	*caviae*		1
GN	*Aeromonas*	*caviae-hydrophila*		12
GN	*Aeromonas*	*hydrophila*		1
GN	*Aeromonas*	*sobria-veronii*		9
GN	*Aureimonas*	*altamirensis*		1
GN	*Bacteroides*	*fragilis*		2
GN	*Brevundimonas*	*diminuta*		1
GN	*Brevundimonas*	*species*		1
GN	*Burkholderia*	*cepacia*		55
GN	*Burkholderia*	*gladioli*		16
GN	*Burkholderia*	*multivorans*		1
GN	*Burkholderia*	*species*		30
GN	*Chryseobacterium*	*indologenes*		2
GN	*Chryseobacterium*	*meningosepticum*		1
GN	*Chryseobacterium*	*species*		1
GN	*Citrobacter*	*amalonaticus*		5
GN	*Citrobacter*	*braakii*		15
GN	*Citrobacter*	*farmeri*		1
GN	*Citrobacter*	*freundii*		54
GN	*Citrobacter*	*koseri*		6
GN	*Citrobacter*	*sedlakii*		1
GN	*Citrobacter*	*species*		26
GN	*Citrobacter*	*youngae*		4
GN	*Cupriavidus*	*species*		2
GN	*Delftia*	*acidovorans*		1
GN	*Edwardsiella*	*tarda*		1
GN	*Enterobacter*	*aerogenes*		15
GN	*Enterobacter*	*cloacae complex*		44
GN	*Enterobacter*	*hormaechei*		3
GN	*Enterobacter*	*kobei*		25
GN	*Enterobacter*	*species*		54
GN	*Erwinia*	*rhapontici*		1

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
GN	*Escherichia*	*coli*	O157H7	94
GN	*Escherichia*	*coli*		595
GN	*Escherichia*	*fergusonii*		4
GN	*Escherichia*	*hermannii*		1
GN	*Escherichia*	*species*		29
GN	*Escherichia*	*vulneris*		2
GN	*Haemophilus*	*influenzae*		4
GN	*Hafnia*	*alvei*		29
GN	*Hafnia*	*species*		14
GN	*Kingella*	*denitrificans*		1
GN	*Kingella*	*kingae*		23
GN	*Klebsiella*	*aerogenes*		1
GN	*Klebsiella*	*oxytoca*		20
GN	*Klebsiella*	*pneumoniae*		82
GN	*Klebsiella*	*species*		20
GN	*Moraxella*	*catarrhalis*		4
GN	*Morganella*	*morganii*		15
GN	*Neisseria*	*gonorrhoeae*		20
GN	*Ochrobactrum*	*anthropi*		9
GN	*Pantoea*	*agglomerans*		3
GN	*Pantoea*	*species*		7
GN	*Pasteurella*	*canis*		1
GN	*Pasteurella*	*multocida*		1
GN	*Pasteurella*	*species*		1
GN	*Plesiomonas*	*shigelloides*		4
GN	*Proteus*	*hauseri*		1
GN	*Proteus*	*mirabilis*		15
GN	*Proteus*	*pennei*		1
GN	*Proteus*	*vulgaris*		5
GN	*Providencia*	*rettgeri*		2
GN	*Providencia*	*stuartii*		2
GN	*Pseudomonas*	*aeruginosa*		282
GN	*Pseudomonas*	*alcaligenes*		1
GN	*Pseudomonas*	*chlororaphis*		1
GN	*Pseudomonas*	*fluorescens*		9
GN	*Pseudomonas*	*fragi*		2
GN	*Pseudomonas*	*luteola*		1
GN	*Pseudomonas*	*mendocina*		2
GN	*Pseudomonas*	*putida*		8
GN	*Pseudomonas*	*stutzeri*		1
GN	*Rahnella*	*aquatilis*		1

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
GN	*Raoultella*	*ornithinolytica*		3
GN	*Raoultella*	*planticola*		1
GN	*Roseomonas*	*genomospecies*		1
GN	*Salmonella*	*bongori*	Brookfield	2
GN	*Salmonella*	*bongori*		3
GN	*Salmonella*	*enterica*	Abaetetuba	1
GN	*Salmonella*	*enterica*	Abony	1
GN	*Salmonella*	*enterica*	Adelaide	1
GN	*Salmonella*	*enterica*	Agona	21
GN	*Salmonella*	*enterica*	Alachua	1
GN	*Salmonella*	*enterica*	Amsterdam	1
GN	*Salmonella*	*enterica*	Anatum	2
GN	*Salmonella*	*enterica*	Antum	2
GN	*Salmonella*	*enterica*	Arizonae	1
GN	*Salmonella*	*enterica*	Babelsberg	1
GN	*Salmonella*	*enterica*	Bareilly	1
GN	*Salmonella*	*enterica*	Bleadon	1
GN	*Salmonella*	*enterica*	Bovismorbificans	4
GN	*Salmonella*	*enterica*	Braenderup	13
GN	*Salmonella*	*enterica*	Cerro	1
GN	*Salmonella*	*enterica*	Chester	1
GN	*Salmonella*	*enterica*	cholera-suis	1
GN	*Salmonella*	*enterica*	Cubana	13
GN	*Salmonella*	*enterica*	Durban	1
GN	*Salmonella*	*enterica*	Ealing	1
GN	*Salmonella*	*enterica*	Enteritidis	43
GN	*Salmonella*	*enterica*	Freetown	1
GN	*Salmonella*	*enterica*	Gaminara	1
GN	*Salmonella*	*enterica*	groupB	20
GN	*Salmonella*	*enterica*	groupC1	6
GN	*Salmonella*	*enterica*	groupC2	2
GN	*Salmonella*	*enterica*	groupC2-C3	3
GN	*Salmonella*	*enterica*	groupC3	2
GN	*Salmonella*	*enterica*	groupD	12
GN	*Salmonella*	*enterica*	groupE	2
GN	*Salmonella*	*enterica*	Hadar	73
GN	*Salmonella*	*enterica*	Halle	1
GN	*Salmonella*	*enterica*	Hartford	1
GN	*Salmonella*	*enterica*	Havana	10
GN	*Salmonella*	*enterica*	Heidelberg	94
GN	*Salmonella*	*enterica*	Infantis	12

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
GN	*Salmonella*	*enterica*	Johannesburg	7
GN	*Salmonella*	*enterica*	Kentucky	60
GN	*Salmonella*	*enterica*	Kiambu	2
GN	*Salmonella*	*enterica*	L-8,20-i-	1
GN	*Salmonella*	*enterica*	Larochelle	1
GN	*Salmonella*	*enterica*	liib-61-k-1.5	1
GN	*Salmonella*	*enterica*	Lille	5
GN	*Salmonella*	*enterica*	Livingstone	12
GN	*Salmonella*	*enterica*	London	3
GN	*Salmonella*	*enterica*	Mbandaka	1
GN	*Salmonella*	*enterica*	Meleagridis	1
GN	*Salmonella*	*enterica*	Minnesota	1
GN	*Salmonella*	*enterica*	Mishmar	1
GN	*Salmonella*	*enterica*	Mishmar-Haemek	1
GN	*Salmonella*	*enterica*	Molade	10
GN	*Salmonella*	*enterica*	Montevideo	10
GN	*Salmonella*	*enterica*	Muenchen	9
GN	*Salmonella*	*enterica*	Ndolo	1
GN	*Salmonella*	*enterica*	Newport	67
GN	*Salmonella*	*enterica*	Ohio	16
GN	*Salmonella*	*enterica*	Oranienburg	7
GN	*Salmonella*	*enterica*	Orion	6
GN	*Salmonella*	*enterica*	Othmarschen	1
GN	*Salmonella*	*enterica*	Panama	9
GN	*Salmonella*	*enterica*	Paratyphi	1
GN	*Salmonella*	*enterica*	Pollorum	1
GN	*Salmonella*	*enterica*	Pomona	5
GN	*Salmonella*	*enterica*	Poona	1
GN	*Salmonella*	*enterica*	Putten	1
GN	*Salmonella*	*enterica*	Rissen	1
GN	*Salmonella*	*enterica*	Saintpaul	3
GN	*Salmonella*	*enterica*	Sandiego	1
GN	*Salmonella*	*enterica*	Schwarzengrund	34
GN	*Salmonella*	*enterica*	Senftenberg	17
GN	*Salmonella*	*enterica*	Soerenga	1
GN	*Salmonella*	*enterica*	Tennessee	1
GN	*Salmonella*	*enterica*	Thomasville	1
GN	*Salmonella*	*enterica*	Thompson	59
GN	*Salmonella*	*enterica*	Tumodi	1
GN	*Salmonella*	*enterica*	Typhimurium	120
GN	*Salmonella*	*enterica*	Vom	1

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
GN	*Salmonella*	*enterica*	Zwickau	1
GN	*Salmonella*	*enterica*		87
GN	*Serratia*	*fonticola*		1
GN	*Serratia*	*marcescens*		32
GN	*Shigella*	*boydii*		2
GN	*Shigella*	*dysenteriae*		2
GN	*Shigella*	*flexneri*		70
GN	*Shigella*	*sonnei*		80
GN	*Shigella*	*species*		6
GN	*Stenotrophomonas*	*species*		55
GN	*Vibrio*	*cholerae*		2
GN	*Vibrio*	*fluvialis*		1
GN	*Vibrio*	*parahaemolyticus*		7
GN	*Vibrio*	*vulnificus*		1
GN	*Yersinia*	*enterocolitica*		2
GN	*Yersinia*	*frederiksenii*		3
GN	*Yersinia*	*intermedia*		3
GP	*Aerococcus*	*urinae*		2
GP	*Bacillus*	*amyloliquefaciens*		1
GP	*Bacillus*	*cereus*		24
GP	*Bacillus*	*circulans*		2
GP	*Bacillus*	*lichenformis*		1
GP	*Bacillus*	*megaterium*		11
GP	*Bacillus*	*species*		18
GP	*Bacillus*	*subtilis*		2
GP	*Bacillus*	*thuringiensis*		6
GP	*Clostridium*	*difficile*		19
GP	*Corynebacterium*	*amycolatum*		2
GP	*Corynebacterium*	*aurimucosum*		3
GP	*Corynebacterium*	*bovis*		1
GP	*Corynebacterium*	*ihumii*		2
GP	*Corynebacterium*	*imitans*		1
GP	*Corynebacterium*	*jeikeium*		2
GP	*Corynebacterium*	*riegelii*		1
GP	*Corynebacterium*	*species*		2
GP	*Corynebacterium*	*striatum*		4
GP	*Corynebacterium*	*tuberculostearicum*		1
GP	*Dermabacter*	*hominis*		4
GP	*Enterococcus*	*avium*		1
GP	*Enterococcus*	*casseliflavis*		8
GP	*Enterococcus*	*faecalis*	VRE	21

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
GP	*Enterococcus*	*faecalis*	VSE	20
GP	*Enterococcus*	*faecalis*		47
GP	*Enterococcus*	*faecium*	VRE	220
GP	*Enterococcus*	*faecium*	VSE	24
GP	*Enterococcus*	*faecium*		4
GP	*Enterococcus*	*gallinarum*		3
GP	*Enterococcus*	*species*	VRE	11
GP	*Gemella*	*haemolysans*		1
GP	*Gordonia*	*bronchialis*		1
GP	*Kocuria*	*kristinae*		2
GP	*Kocuria*	*marina*		1
GP	*Kocuria*	*rhizophila*		1
GP	*Kytococcus*	*sedentarius*		1
GP	*Lactobacillus*	*fermentum*		1
GP	*Lactobacillus*	*species*		3
GP	*Listeria*	*grayi*		4
GP	*Listeria*	*innocua*		9
GP	*Listeria*	*ivanovii*		6
GP	*Listeria*	*monocytogenes*		89
GP	*Listeria*	*murrayi*		1
GP	*Listeria*	*seeligeri*		6
GP	*Listeria*	*welshimeri*		12
GP	*Micrococcus*	*luteus*		16
GP	*Micrococcus*	*lylae*		2
GP	*Micrococcus*	*species*		1
GP	*Micrococcus*	*yunnanensis*		1
GP	*Paracoccus*	*yeei*		2
GP	*Rhodococcus*	*equi*		1
GP	*Rothia*	*amarae*		1
GP	*Rothia*	*mucilaginosa*		3
GP	*Staphylococcus*	*aureus*	CMRSA	102
GP	*Staphylococcus*	*aureus*	MRSA	176
GP	*Staphylococcus*	*aureus*	MSSA	330
GP	*Staphylococcus*	*aureus*		458
GP	*Staphylococcus*	*capitis*		33
GP	*Staphylococcus*	*caprae*		10
GP	*Staphylococcus*	*carnosus*		1
GP	*Staphylococcus*	*caseolyticus*		1
GP	*Staphylococcus*	*cohnii*		4
GP	*Staphylococcus*	*epidermidis*		16
GP	*Staphylococcus*	*equorum*		1

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
GP	*Staphylococcus*	*haemolyticus*		19
GP	*Staphylococcus*	*hominis*		41
GP	*Staphylococcus*	*lugdunensis*		18
GP	*Staphylococcus*	*pasteuri*		9
GP	*Staphylococcus*	*saccharolyticus*		3
GP	*Staphylococcus*	*sciuri*		1
GP	*Staphylococcus*	*simulans*		6
GP	*Staphylococcus*	*species*		18
GP	*Staphylococcus*	*vitulinus*		1
GP	*Staphylococcus*	*warneri*		24
GP	*Staphylococcus*	*xylosus*		2
GP	*Streptococcus*	*agalactiae*		12
GP	*Streptococcus*	*anginosus*		15
GP	*Streptococcus*	*canis*		3
GP	*Streptococcus*	*constellasus*		10
GP	*Streptococcus*	*cristatus*		2
GP	*Streptococcus*	*dysgalactiae*		5
GP	*Streptococcus*	*gallolyticus*		2
GP	*Streptococcus*	*gordonii*		2
GP	*Streptococcus*	*groupA*		13
GP	*Streptococcus*	*groupB*		12
GP	*Streptococcus*	*groupC*		4
GP	*Streptococcus*	*groupG*		10
GP	*Streptococcus*	*infantarius*		3
GP	*Streptococcus*	*intermedius*		5
GP	*Streptococcus*	*mitis*		5
GP	*Streptococcus*	*mitis-group*		1
GP	*Streptococcus*	*mitis-oralis*		10
GP	*Streptococcus*	*mutans*		2
GP	*Streptococcus*	*parasanguinis*		10
GP	*Streptococcus*	*pneumoniae*		20
GP	*Streptococcus*	*pyogenes*		14
GP	*Streptococcus*	*salivarius*		17
GP	*Streptococcus*	*sanguinis*		6
GP	*Streptococcus*	*species*		6
GP	*Streptococcus*	*viridans*		6
GP	*Streptococcus*	*viridans-group*		2
GP	*Streptomyces*	*carnosus*		1
GP	*Turicella*	*otitidis*		2
MD	*Penicillium*	*roqueforti*		9
YT	*Candida*	*albicans*		171

Continued

Gram-type	Genus	Species	Strain	No. isolates collected
YT	*Candida*	*auris*		20
YT	*Candida*	*bracarensis*		1
YT	*Candida*	*dubliniensis*		12
YT	*Candida*	*duobushaemulonii*		6
YT	*Candida*	*famata*		1
YT	*Candida*	*glabrata*		70
YT	*Candida*	*guilliermondii*		12
YT	*Candida*	*haemulonii*		4
YT	*Candida*	*inconspicua*		1
YT	*Candida*	*kefyr*		9
YT	*Candida*	*krusei*		17
YT	*Candida*	*lambica*		1
YT	*Candida*	*lipolytica*		1
YT	*Candida*	*lusitaniae*		16
YT	*Candida*	*norvegensis*		1
YT	*Candida*	*orthopsilosis*		7
YT	*Candida*	*palmioleophila*		2
YT	*Candida*	*parapsilosis*		69
YT	*Candida*	*pararugosa*		3
YT	*Candida*	*pelliculosa*		1
YT	*Candida*	*rugosa*		1
YT	*Candida*	*species*		48
YT	*Candida*	*sphaerica*		1
YT	*Candida*	*tropicalis*		22
YT	*Candida*	*utilis*		3
YT	*Candida*	*zeylanoides*		1
YT	*Cryptococcus*	*albidus*		6
YT	*Cryptococcus*	*diffluens*		23
YT	*Cryptococcus*	*gatti*		8
YT	*Cryptococcus*	*laurentii*		6
YT	*Cryptococcus*	*magnus*		1
YT	*Cryptococcus*	*neoformans*		17
YT	*Cryptococcus*	*terreus*		1
YT	*Cryptococcus*	*uniguttulatus*		4
YT	*Exophiala*	*dermatitidis*		9
YT	*Geotrichum*	*candidum*		1
YT	*Malassezia*	*furfur*		1
YT	*Malassezia*	*pachydermatis*		1
YT	*Meyerozyma*	*caribbica*		1
YT	*Prototheca*	*wickerhamii*		2
YT	*Rhodotorula*	*glutinis*		2

Continued				
Gram-type	**Genus**	**Species**	**Strain**	**No. isolates collected**
YT	*Rhodotorula*	*minuta*		1
YT	*Rhodotorula*	*mucilaginosa*		4
YT	*Saccharomyces*	*cerevisiae*		16
YT	*Saprochaete*	*capitata*		1
YT	*Sporobolomyces*	*salmonicolor*		1
YT	*Trichosporon*	*asahii*		4
YT	*Trichosporon*	*dermatis*		1
YT	*Trichosporon*	*faecale*		1
YT	*Trichosporon*	*inkin*		1
YT	*Trichosporon*	*mucoides*		2
YT	*Trichosporon*	*mycotoxinivorans*		1

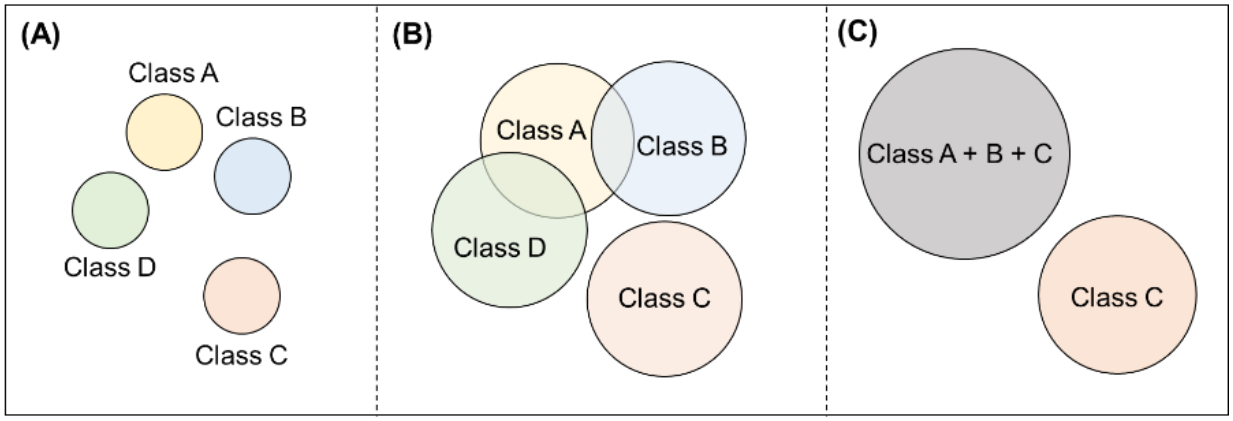


Figure A.4. Visual representation of (A) complete separation of four groups of spectra (Class A, B, C and D) while (B) represents the same groups in (A), however, there is an increase in representation of each class and (C) represents grouping of Classes A, B and C to obtain a well-separated pair of the increased in representation classes

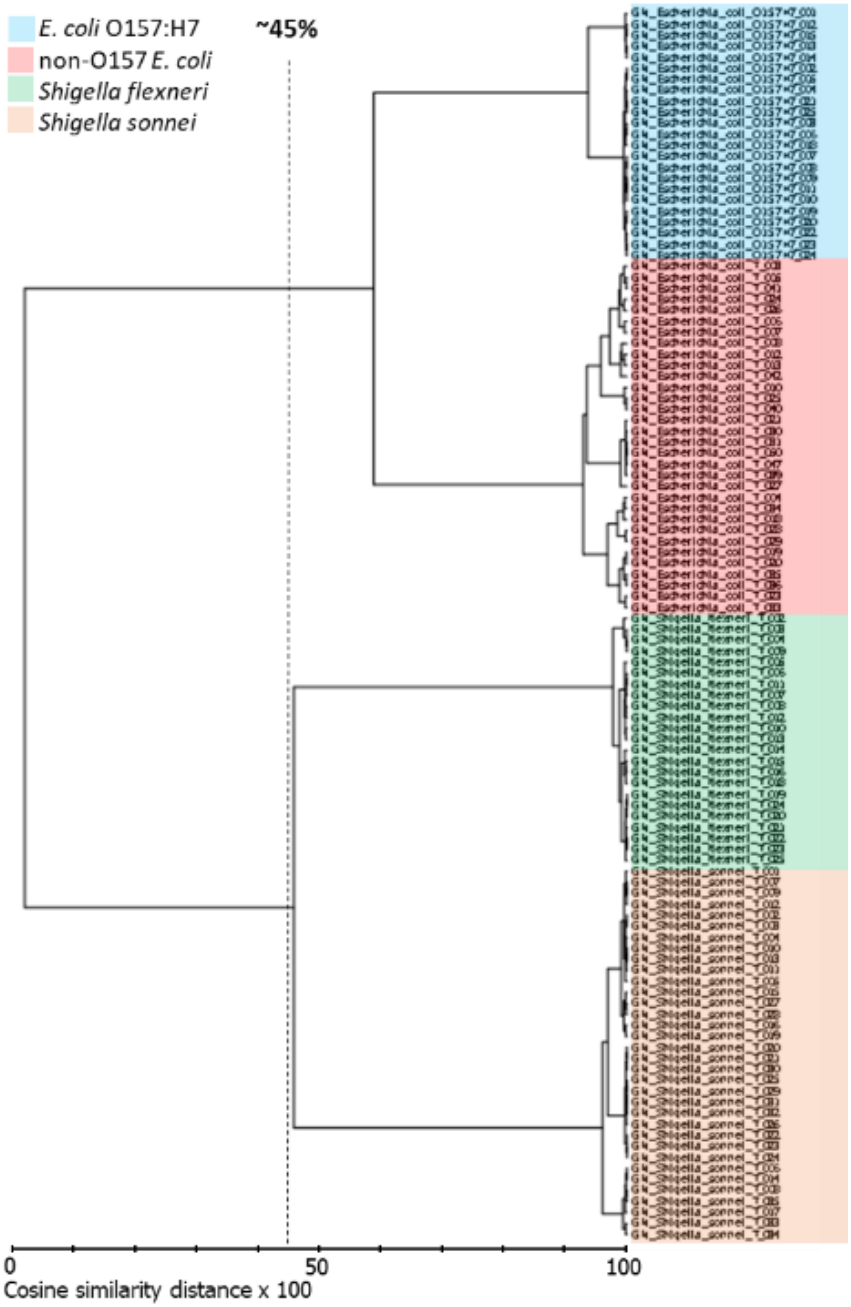


Figure A.5. Dendrogram generated from hierarchical cluster analysis and feature selection of ATR-FTIR spectra of non-O157 *Escherichia coli*, *E. coli* O157:H7, *Shigella flexneri* and *S. sonnei* depicting 45% spectral similarity between the two genera

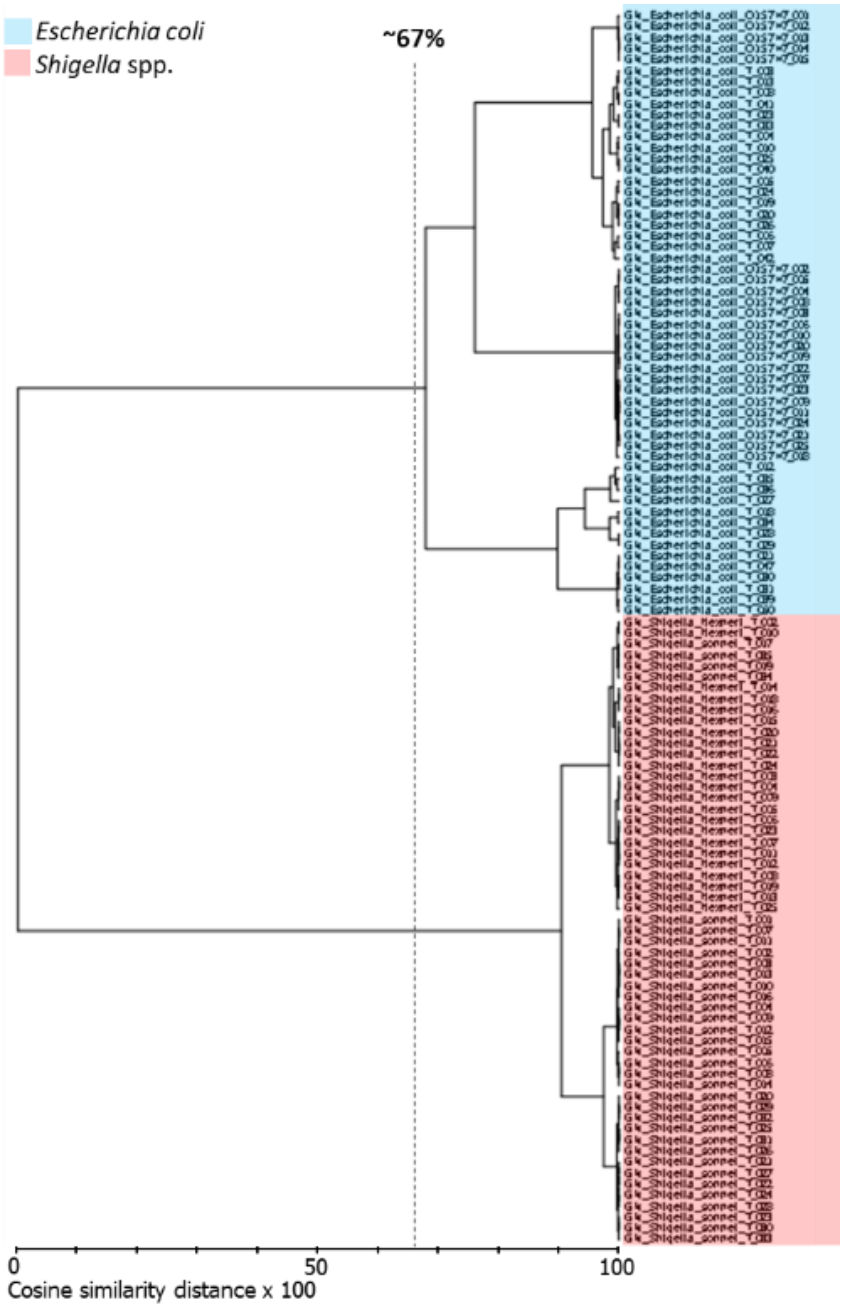


Figure A.6. Dendrogram generated from hierarchical cluster analysis and feature selection of ATR-FTIR spectra of *Escherichia coli* and *Shigella* species depicting 67% spectral similarity between the two genera

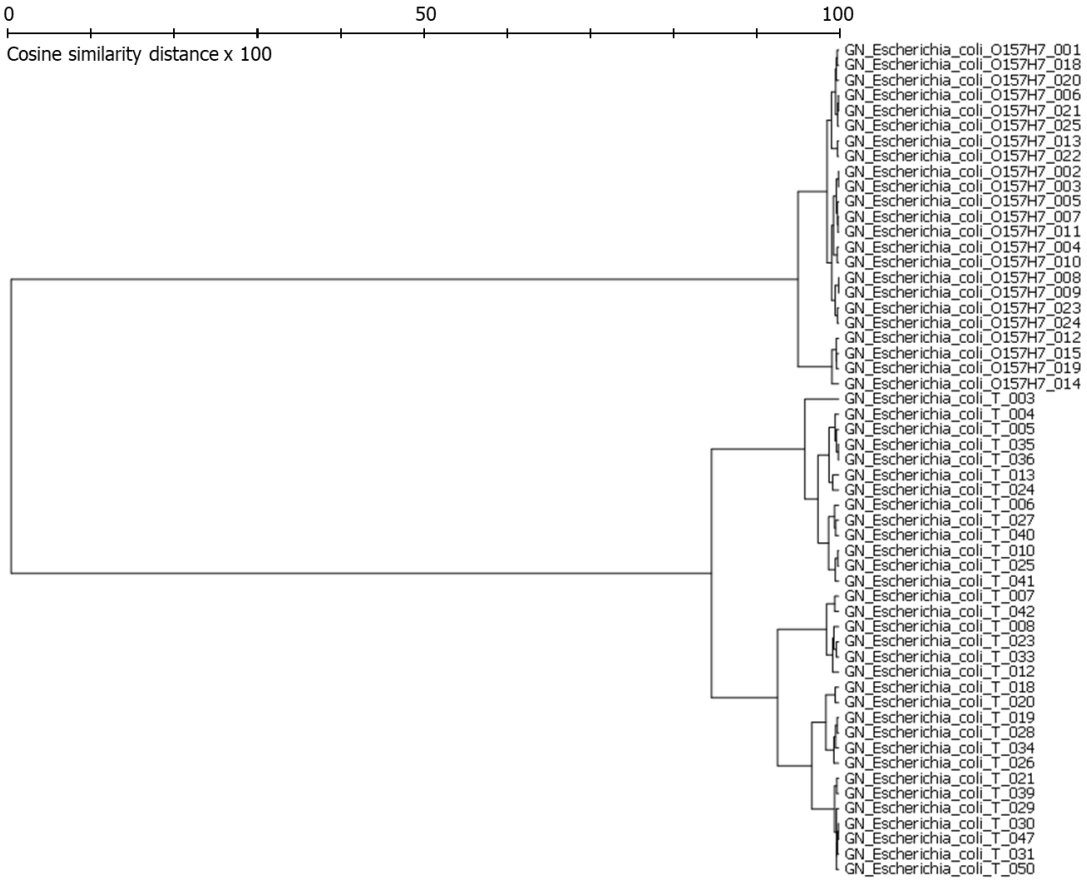


Figure A.7. Dendrogram generated from hierarchical cluster analysis and feature selection of ATR-FTIR spectra of *Escherichia coli* O157:H7 and non-O157 *E. coli*

Table A.7. Clinical microbiology laboratories and isolation sources of isolates collected in CHAPTER 5

Code	Clinical microbiology laboratory	Code	Isolation source
A	Centre de SSS D'Arthabaska-et-de-l'érable Hôtel-Dieu	a	Abscess
B	Centre de SSS de Trois-Rivières	b	Aspiration
C	Centre hospitalier régional du Grand Portage	c	Tissue
D	CHU Sainte-Justine	d	Bile
E	CSSS Haut-Richelieu/Rouville (Hôpital)	e	Bronchoscopy
F	Hôpital Charles-LeMoyne	f	Tonsil swab
G	Hôpital de Gatineau	g	Caterer
H	Hôpital de Hull	j	Drain
I	Hôpital de Rouyn-Noranda	i	Blood
J	Hôpital du Centre-de-la-Mauricie	j	Bronchial wash
K	Hôpital du Sacré-Cœur de Montréal	k	Ascite fluid
L	Hôpital Fleurimont	l	Lung fluid
M	Hôpital général du Lakeshore	m	Gastric fluid
N	Hôpital Honoré-Mercier	n	Pus
O	Hôpital Pierre-Boucher	o	Wound
P	Hôpital régional de Rimouski	p	Secretion
Q	Hôpital régional de Saint-Jérôme	q	Urine
R	Hôpital Saint-Luc du CHUM - Microbiologie	r	Phlem/mucus
S	Hôpital Sainte-Croix	s	Secretion
T	Laboratoire médical Biron inc.	t	Skin
U	Laboratoires médicaux CDL inc.	u	Other
V	Pavillon Hôtel-Dieu		
W	Hôpital Santa Cabrini		
X	Site Glen - LAB Microbiologie		
Y	Centre de SSS de la Haute-Yamaska		
Z	Centre hospitalier Anna-Laberge		
AA	Hôpital Brôme-Missisquoi-Perkins		
BB	Hôpital du Suroît		
CC	Inst. Univ. de cardiologie et de pneumologie de QC		
DD	Centre Hospitalier Régional de Lanaudière		
EE	Hôpital Maisonneuve-Rosemont		
FF	Hôpital de Verdun		
GG	Hôpital Notre-Dame du CHUM		
HH	Hôpital de l'Enfant-Jésus		
II	Hôpital de Papineau		
JJ	Hôpital de Chicoutimi		
KK	Hôpital régional de Rimouski		
LL	Hôpital général Juif		

Table A.8. Tabulation of Source and origin of routine clinical isolates collected in CHAPTER 5 (refer to Table A.7 for letter codes associated with clinical microbiology laboratories and isolation source)

Microorganism	No. isolates	No. of different clinical microbiology laboratory (clinical microbiology laboratory)	No. of different isolation source (Isolation source)
Candida spp.			
C. albicans	130	22(A-V)	17(a-q)
C. dubliniensis	13	5(G, M, R, W, X)	6(d, j, l, n, q, u)
C. glabrata	68	18(B, D-I, K, L, N, R, V, X-CC)	13(a-c, e, g, j, k, l, n, o, r, q, u)
C. guilliermondii	1	1(Q)	1(u)
C. kefyr	2	1(X	1(g)
C. krusei	8	6(D, R, X, CC-EE)	4(u, r, l, q)
C. lipolytica	1		
C. lusitaniae	16	4(R, X, GG, C)	6(a, b, n, q, r, u)
C. orthopsilosis	4	4(D, M, R, X)	3(i, q, r)
C. parapsilosis	32	12(D, G, H, J, L, M, R, V, CC, EE, HH, II)	8(c, e, l, n, o, q, t, u)
C. pararugosa	1	1(X)	1(c)
C. pelliculosa	1	1(JJ)	1(i)
C. tropicalis	23	8(D, H, L, R, BB, CC, GG, KK)	8(a, c, e, l, n, p, q, u)
C. utilis	3	2(D, R)	1(q)
Cryptococcus neoformans	7	7(D, K, L, N, R, X, DD)	2(c, j)
Meyerozyma caribbica	1	1(D)	1(a)
Rhodotorula mucilaginosa	1	1(LL)	1(u)
Saccharomyces cerevisiae	5	4(F, R, W, LL)	4(a, l, s, u)
Trichosporon spp.	1	1(CC)	1(r)

Table A.9. Complete list of species represented in the ATR-FTIR reference spectral database by microorganism name and total isolates present

Microorganism	Total	Microorganism	Total
Candida spp.		*Malassezia* spp.	
C. albicans	16	*M. globosa*	1
C. auris	11	*M. furfur*	1
C. bracarensis (*Nakaseomyces bracarensisa*)	1	*M. pachydermatis*	2
C. dubliniensis	13	*M. sympodialis*	1
C. duobushaemulonii	3	*M. slooffiae*	1
C. ciferrii	1	*Meyerozyma* spp.	
C. (*Nakaseomyces*) *glabrata*	17	*M. caribbica* (*Candida fermentati*)	5
C. inconspicua (Pichia cactophila)	1	*M.* (*Candida*) *guilliermondii*	15
C. intermedia	1	*Naganishia* spp.	
C. metapsilosis	1	*N. albida* (*Cryptococcus albidus*)	3
C. melibiosica	1	*N.* (*Cryptococcus*) *diffluens*	2
C. nivariensis (*Nakaseomyces nivariensisa*)	1	*Pichia* spp.	
C. orthopsilosis	9	*P. fermentans* (*Candida lambica*)	1
C. parapsilosis	12	*P. kudriavzevii* (*Candida krusei*)	10
C. palmioleophila	1	*P. membranifaciens* (*Candida valida*)	1
C. (Metschnikowia) pulcherrima	1	*P.* (*Candida*) *norvegensis*	1
C. (Starmerella) sorbosivorans	1	*Prototheca wickerhamii*	2
C. tropicalis	15	*Rhodotorula* spp.	
C. viswanathii	1	*R. mucilaginosa*	3
C. zeylanoides	1	*R. glutinis*	1
Clavispora (*Candida*) *lusitaniae*	10	*R. minuta*	1
Cryptococcus spp.		*Saccharomyces cerevisiae*	13
C. gattii	5	*Saprochaete clavata* (*Geotrichum clavatum*)	1
C. (Papiliotrema) laurentii	3	*Sporobolomyces salmonicolor*	1
C. (Filobasidium) magnus	1	*Starmerella* (*Candida*) *magnoliae*	1
C. neoformans	9	*Trichosporon* spp.	
C. terreus (Solicoccozyma terrea)	1	*T. asahii*	4
C. uniguttulatus	1	*T. (Cutaneotrichosporon) mucoides*	2
Cyberlindnera jadinii (*Candida utilis*)	4	*T. (Cutaneotrichosporon) dermatis*	1
Debaryomyces hansenii (*Candida famata*)	1	*T. faecale*	1
Diutinia spp.		*T. inkin*	1
D. (*Candida*) *catenulata*	1	*T.* (*Apiotrichum*) *mycotoxinivorans*	1
D. (*Candida*) *rugosa*	1	*Wickerhamiella* (*Candida*) *pararugosa*	4
Exophiala dermatitidis	9	*Wickerhamomyces anomalus* (*Candida pelliculosa*)	2
Geotrichum candidum	1	*Yarrowia* (*Candida*) *lipolytica*	7
Kluyveromyces spp.			
K. marxianus (*Candida kefyr*)	17		
K. lactis (*Candida sphaerica*)	1		

Table A.10. Additional isolates collected at Centre hospitalier Sainte-Justine for the evaluation (retrospective) of the ATR-FTIR spectroscopy-based microorganism identification technique

Microorganism	**Isolates collected**	**Correct species identification (%)**
Candida		
C. albicans	12	12
C. (*Nakaseomyces*) *glabrata*	4	4
C. (Clavispora) lusitaniae	7	7
C. orthopsilosis	1	1
C. parapsilosis	11	11
C. tropicalis	1	1
Kluyveromyces marxianus (*Candida kefyr*)	1	1
Pichia kudriavzevii (*Candida krusei*)	2	2
Total	**39**	**39**

Table A.11. Detailed information on Chilean outbreak strains of *Candida parapsilosis* (hospital, center, infection date and halotype through microsatellite strain typing)

MY### ID	Hospital	Center	Infection-date	HALOTYPE
MY076471	Aguirre	Aguirre	2011-10-07	AB[AC]E[AC]
MY076472	Aguirre	Aguirre	2012-05-04	AB[AC]E[AC]
MY076473	Aguirre	Aguirre	NA	CBDEA
MY076474	Juan de Dios	Juan de Dios	2012-08-09	BA[AB]D[DE]
MY076475	Aguirre	Aguirre	2012-04-11	AB[AC]E[AC]
MY076476	Juan de Dios	Cerro Navia	2012-07-19	AB[AC]E[AC]
MY076477	Juan de Dios	Juan de Dios	2012-08-09	AB[AC]E[AC]
MY076478	Aguirre	Aguirre	2012-05-05	AB[AC]E[AC]
MY076479	Juan de Dios	Mendoza	2012-07-03	BC[AE]CC
MY076480	San Jose	Aguirre	2012-04-10	AB[AC]E[AC]
MY076481	Juan de Dios	Juan de Dios	2012-07-31	BC[AE]DC
MY076482	H.FACH	Alfa Dial	2012-07-27	BA[AB]A[DE]
MY076483	San Jose	Saint Joseph	2011-09-16	AB[AC]EB
MY076484	San Jose	Dialisis Norte	2012-06-30	DAAFC
MY076485	Aguirre	Aguirre	2011-04-26	AB[AC]E[AD]
MY076486	Aguirre	Aguirre	2012-05-04	BAAAB
MY076487	Aguirre	Aguirre	2012-06-26	BA[AC]DC
MY076488	Aguirre	Aguirre	2010-05-05	ABAE[D]
MY076489	San Jose	Aguirre	2012-05-01	AB[AC]E[AC]
MY076490	C Indisa	La Pintana	2012-08-04	AB[AC]E[AC]
MY076491	San Jose	San Jose	2012-05-19	AB[AC]E[AC]
MY076492	San Jose	Dialisis Norte	2012-06-11	DAAFC
MY077951	San Borja	San Borja	2013-01-14	ABAB[BC]
MY077952	San Borja	San Borja	2012-11-08	BA[AB]C[DE]
MY077953	Pto Montt	Pto Montt	2013-03-08	AB[AC][DE]C
MY077954	San Borja	San Borja	2013-03-09	BA[AB]C[DE]
MY077955	Hurtado	San Gabriel	2013-02-13	AB[AC]E[AC]
MY077956	Hurtado	CD3	2012-12-06	ABA[AB]A
MY077957	San Borja	San Borja	2012-12-31	AB[AC]EC
MY077958	Hurtado	Rosita	2013-03-05	ABAB[BC]
MY077959	San Borja	San Borja	2013-01-25	ABAB[BC]
MY077960	Hurtado	San Gabriel	2013-03-06	ABCE[AC]
MY077961	Hurtado	La Serena	2013-01-25	ABA[AB]A
MY077962	Pto Montt	Pto Montt	2013-04-30	AB[AC]EC
MY077963	HCUCH	CDQ	2013-03-06	-
MY077964	HCUCH	CDQ	2013-03-05	ABA[DE]C
MY077965	Hurtado	CD4	2013-01-04	AB[AC]E[AC]
MY077966	Hurtado	CD4	2013-01-14	AB[AC]EA
MY077967	Hurtado	CD4	2013-01-21	AB[AC][DE][AC]
MY077968	Hurtado	CD4	2012-12-07	AB[AC]EC
MY077969	Hurtado	San Gabriel	2013-03-04	AB[AC]E[AC]
MY077970	Hurtado	San Gabriel	2013-02-20	AB[AC]E[AC]

Table A.12. Concentration in µg/mL of dehydrated antifungal in a 96-well microtiter plate for antimicrobial susceptibility testing

Antifungal		5FC[9]											
Antifungal		1	2	3	4	5	6	7	8	9	10	11	12
AMP[1]	A	2	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8	16
ANI[2]	B	4	0.008	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8
CAS[3]	C	8	0.008	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8
MIC[4]	D	16	0.008	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8
FLU[5]	E	32	0.12	0.25	0.5	1	2	4	8	16	32	64	128
ITR[6]	F	0	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8	16
POS[7]	G	0	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8	16
VOR[8]	H	0	0.016	0.03	0.06	0.12	0.25	0.5	1	2	4	8	16

[1]Amphotericin B, [2]Anidulafungin, [3]Caspofungin, [4]Micafungin, [5]Fluconazole, [6]Itraconazole, [7]Posaconazole, [8]Voriconazole, [9]Flucytosine

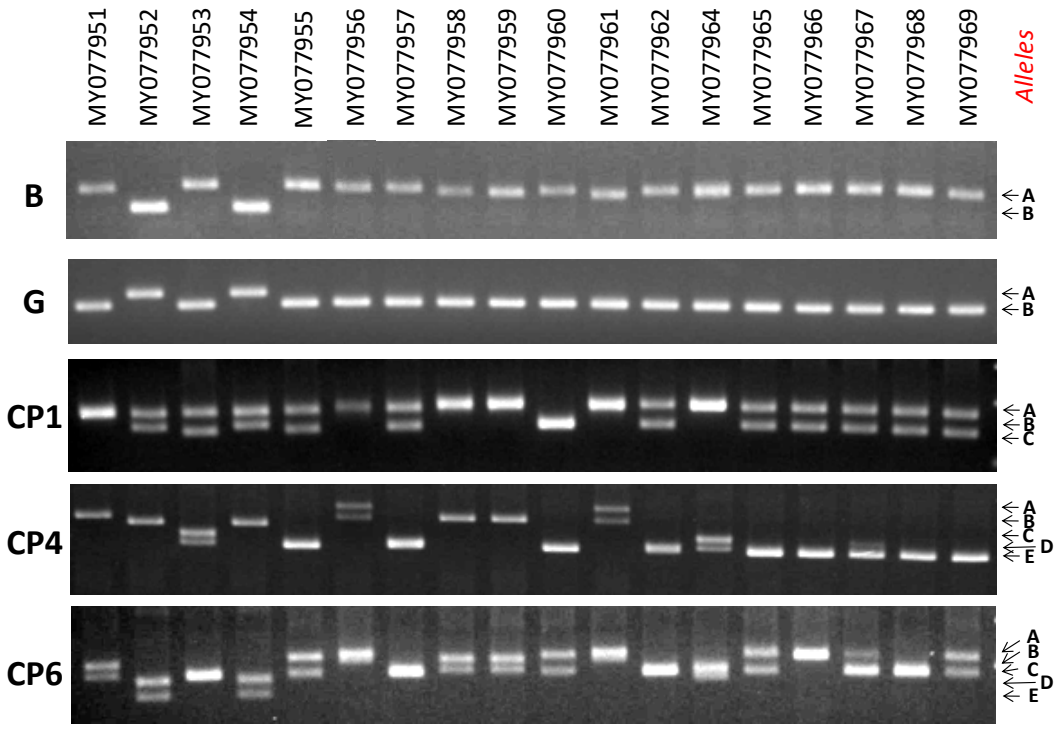


Figure A.8. Example of the gel electrophoresis profile obtained with B, G, CP1α, CP4α and CP6α microsatellites on the C. *parapsilosis* isolates received in 2013

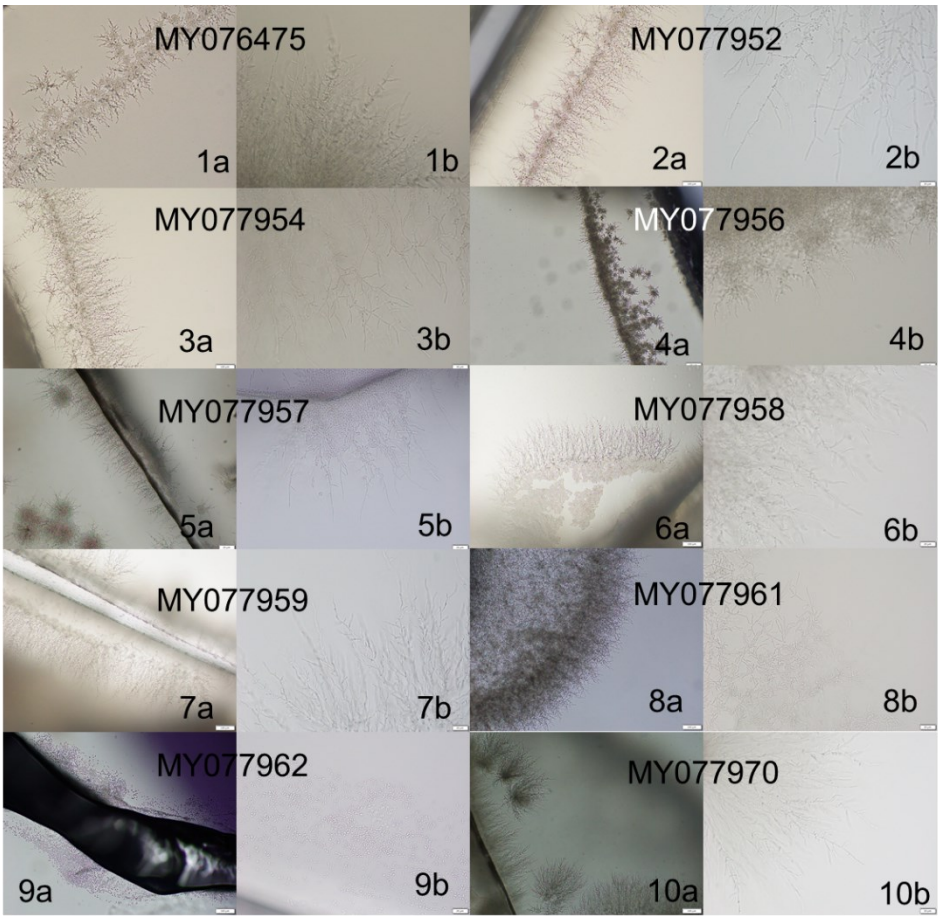


Figure A.9. Pictures of 10 potential Chilean *Candida parapsilosis* outbreak isolates grown on cornmeal agar captured at (a) 10x and (b) 40x

Table A.13. Good coverage for accurate variant calling and number of reads relating the microsatellite molecular typing (MSMT) clusters (chromosome length 14618400 bp)

MSMT cluster	Strain	No. of reads	Coverage
CL5	MY077952	5932630	119
CL5	MY077954	5806980	116
CL2	MY077962	6379038	128
CL2	MY077968	7289986	146
CL1	M4076475	7721802	154
CL3	M4077956	2868561	57
CL4	M4077958	6325474	127
CL4	M4077959	8280804	166
CL3	M4077961	7219860	144
CL1	M4077970	6026720	121

Table A.14. SNP difference matrix between *Candida parapsilosis* isolates from a potential outbreak in Chile based on core single nucleotide variant whole genome sequencing for epidemiological analysis

MST Cluster	CL5		CL5	CL4	CL4	CL2	CL2	CL3	CL3	CL1	CL1
Strain	MY077954	REF	MY077952	MY077959	MY077958	MY077968	MY077962	MY077956	MY077961	MY076475	MY077970
MY077954	0	2381	1	2099	2097	1499	1495	2016	2015	1489	1486
REF	2381	0	2380	1842	1840	2022	2018	1759	1762	2012	2009
MY077952	1	2380	0	2098	2096	1498	1494	2015	2014	1488	1485
MY077952	2099	1842	2098	0	4	1816	1812	471	476	1806	1803
MY077958	2097	1840	2096	4	0	1814	1810	469	474	1804	1801
MY077968	1499	2022	1498	1816	1814	0	34	1731	1734	28	13
MY077962	1495	2018	1494	1812	1810	34	0	1727	1730	18	21
MY077956	2016	1759	2015	471	469	1731	1727	0	31	1721	1718
MY077961	2015	1762	2014	476	474	1734	1730	31	0	1724	1721
MY076475	1489	2012	1488	1806	1804	28	18	1721	1724	0	15
MY077970	1486	2009	1485	1803	1801	13	21	1718	1721	15	0

Table A.15. Example of an inconclusive result by the ATR-FTIR spectroscopy-based method for a *C. dubliniensis* MALDI-TOF MS-identified microorganism. Isolate was passaged multiple times to ensure purity and grown on chromogenic agar

Unknown isolate filename	Genus	Species	Spectral similarity (%)	Standard deviation from the mean
YT_SAB_GLEN_009_ATRC3_20201008_L1	*Candida*	*albicans*	38.5	2.57
YT_SAB_GLEN_009_ATRC3_20201008_L2	*Candida*	*dubliniensis*	88.7	0.604
YT_SAB_GLEN_009_ATRC3_20201008_L3	*Candida*	*albicans*	46.3	2.634
YT_SAB_GLEN_009_ATRC3_20201008_L4	*Candida*	*albicans*	30.0	1.837
YT_SAB_GLEN_009_ATRC3_20201008_L5	*Candida*	*albicans*	22.3	2.579
YT_SAB_GLEN_009_ATRC3_20201008_L6	*Candida*	*dubliniensis*	90.5	0.269
YT_SAB_GLEN_009_ATRC3_20201008_L7	*Candida*	*dubliniensis*	88.9	-0.116
YT_SAB_GLEN_009_ATRC3_20201008_L8	*Candida*	*albicans*	41.0	2.209
YT_SAB_GLEN_009_ATRC3_20201008_L9	*Candida*	*dubliniensis*	91.3	0.269
YT_SAB_GLEN_009_ATRC3_20201008_L10	*Candida*	*dubliniensis*	93.0	0.558