# Bio-Informatics as a Broad-Spectrum Discipline for the Interpretation of Clinical Microbiology Data

**Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Manisha Goyal

aus Saharanpur (Uttar Pradesh), Indien

Tübingen

2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 27.04.2022 |
| Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatter/-in: | Prof. Dr. Andreas Peschel |
| 2. Berichterstatter/-in: | Prof. Dr. Dr. Alex van Belkum |

# Bio-Informatics as a Broad-Spectrum Discipline for the Interpretation of Clinical Microbiology Data

## Doctor of Philosophy (PhD) Thesis

**Eberhard Karls Universität Tübingen**

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von


## Manisha Goyal

aus Saharanpur, India

an Tübingen


2022

Supervised by

**Prof. Andreas Peschel,**

Eberhard Karls University of Tübingen, Germany

And

**Prof. Alex van Belkum,**

BioMérieux SA, France

# Contents

## List of Figures

## List of Tables

**Abbreviations**

| | |
|---|---|
| HTS | High-Throughput Sequencing |
| NGS | Next-Generation Sequencing |
| IVD | In Vitro Diagnostic |
| WGS | Whole Genome Sequencing |
| PCR | Polymerase Chain Reaction |
| qPCR | Quantitative PCR |
| rt-PCR | Reverse Transcriptase PCR |
| ISR | Intergenic Spacer Region |
| DB | Databases |
| MLST | Multi Locus Sequence Typing |
| AMR | Antimicrobial Resistance |
| API | Analytical Profile Index |
| RFLP | Restriction Fragment Length Polymorphism |
| PFGE | Pulsed-Field Gel Electrophoresis |
| RAPD | Random Amplification Of Polymorphic DNA |
| AFLP | Amplified Fragment Length Polymorphism |
| MLVA | Multi Locus Variable Tandem Repeats Amplification |
| RT-qPCR | Reverse Transcriptase Quantitative PCR |
| RT-LAMP | RT Loop Mediated Isothermal Amplification |
| MS | Mass Spectroscopy |
| GC | Gas Chromatography |
| MALDI-TOF | Matrix-Assisted Laser Desorption Ionization Time-Of-Flight |
| CE | Capillary Electrophoresis |
| ESI | Electrospray Ionization |
| SERS | Surface Enhanced Raman Spectroscopy |
| ML | Machine Learning |
| TB | Tuberculosis |
| WHO | World Health Organization |
| MDR | Multi Drug Resistant |
| ONT | Oxford Nanopore Technologies |
| SNP | Single Nucleotide Polymorphism |
| *S. aureus* | *Staphylococcus aureus* |
| MSCRAMMS | Microbial Surface Components Recognizing Adhesive Matrix Molecules |
| SCIN | Staphylococcal Complement Inhibitor |
| IEC | Immune Evasion Cluster |
| KB | Kilo Base |
| *C. difficile* | *Clostridioides difficile* |
| CDI | *C. difficile* Infection |
| RT | Ribotype |
| DBGWAS | De Bruijn Graph-Based Genome Wide Association Studies |
| REA | Restriction Endonuclease Analysis |
| NCBI | National Center For Biotechnology And Information |
| FP | False-Positives |
| FN | False-Negatives |

| | |
|---|---|
| TP | True-Positives |
| TN | True-Negatives |
| FDR | False Discovery Rate |
| cDBG | Compacted De Bruijn Graphs |
| CC | Clonal Complex |
| cgMLST | Core Genome Multi Locus Sequence Typing |
| SARS-CoV-2 | SARS Coronavirus-2 |
| CDS | Coding Sequences |
| RdRp | RNA-Dependent RNA Polymerase |
| GISAID | Global Initiative On Sharing Avian Influenza Data |
| MST | Minimum Spanning Tree |
| CNPA | Carbapenem Non-Susceptible *Pseudomonas aeruginosa* |
| ICUs | Intensive Care Units |
| wgMLST | Whole Genome-Based Multi Locus Sequence Typing |
| wgSNP | Whole Genome SNP |
| *P. aeruginosa* | *Pseudomonas aeruginosa* |
| AST | Antibiotic Susceptibility Testing |
| ROC curve | Receiver Operating Characteristic Curve |
| SNV | Single-Nucleotide Variant |
| HAI | Healthcare-Associated Infections |
| VNTRs | Variable Numbers Of Tandem Repeats |
| RFLP | Restriction Fragment Length Polymorphisms |
| AFLP | Amplified Fragment Length Polymorphisms |
| GUI | Graphical User Interfaces |

**List of publications and Manuscripts**

1. **Goyal, M.;** Javerliat, F.; Palmieri, M.; Mirande, C.; van Wamel, W.; Tavakol, M.; Verkaik, N. J. and van Belkum, A. (2019). Genomic Evolution of Staphylococcus aureus During Artificial and Natural Colonization of the Human Nose. Frontiers in microbiology, 10, 1525. https://doi.org/10.3389/fmicb.2019.01525.
2. **Goyal, M.;** Hauben, L.; Pouseele, H.; Jaillard, M.; De Bruyne, K.; van Belkum, A. and Goering, R. (2020). Retrospective Definition of *Clostridioides difficile* PCR Ribotypes on the Basis of Whole Genome Polymorphisms: A Proof of Principle Study. *Diagnostics* , *10*, 1078. https://doi.org/10.3390/diagnostics10121078
3. **Goyal, M.**; De Bruyne, K.; van Belkum, A.; West, B. (2021). Different SARS-CoV-2 haplotypes associate with geographic origin and case fatality rates of COVID-19 patients, Infection, Genetics and Evolution, 90,104730, ISSN 1567-1348, https://doi.org/10.1016/j.meegid.2021.104730. *(Article is being used for marketing purpose of recently developed plugin tool of BioNumerics for the analysis of SARS-CoV-2)*
4. **Goyal, M.**; Pelegrin, A. C.; Jaillard, M.; Saharman, Y.; Klaassen, C.; Verbrugh, H.; Severin, J.A. and van Belkum, A. (2022). Whole Genome Multi-Locus Sequence Typing And Genomic Single Nucleotide Polymorphism Analysis For Epidemiological Typing Of Pseudomonas Aeruginosa From Indonesian Intensive Care Units. Frontiers Microbiology (Submitted and under review).
5. **Goyal, M**.; Jaillard, M.; Rochas, O. and van Belkum, A. Central Role of Bioinformatics In Next Generation Microbiological Epidemiology And Typing: The Review. (Manuscript under submission)
6. **Goyal, M.;** Moingeon, B.; Bulteau, S.; Dombrecht, J.; De Bruyne, K.; and van Belkum, A. Epidemiological analysis of Pseudomonas aeruginosa using EPISEQ® CS: an advanced one-stop solution for Next Generation Sequencing data analysis. (Manuscript under submission)

# Chapter 1

## Central Role of Bioinformatics In Next Generation Microbiological Epidemiology And Typing: The Review

Manisha Goyal[1], Magali Jaillard[2], Olivier Rochas[3] and Alex van Belkum[1,*]

[1]bioMérieux, Open Innovation and Partnerships, 3 Route du Port Michaud, 38390 La Balme Les Grottes, France
[2]bioMérieux, Data Sciences, Chemin de L'Orme, 69280 Marcy l'Etoile-France
[3]bioMérieux, Corporate Business Development, Chemin de L'Orme, 69280 Marcy l'Etoile-France

**ABSTRACT**

Routine clinical microbiology laboratories are at the receiving end of new diagnostic technologies. Over the past decades nucleic acid amplification technology, mass spectrometry and, more recently, omics technologies including next generation sequencing have been successfully introduced for increasingly large-scale diagnostic applications. Particularly in the field of microbial infections, diagnostics start with an accurate identification of pathogenic microbes. Rapid technological advancements made it efficient through variety of methods from classical to the most recent technologies for strain detection and characterization. The introduction in routine diagnostic laboratories of such technologies in combination with readily available and extensive amounts of clinical patient-related and demographic data has led to a surge in analytical tools for the joined interpretation of both diagnostic laboratory data and patient-oriented information. We will here summarize the diverse methods that are available for interpretation of such large scale diagnostic data and we will summarize the quality of additional tools that will allow the combined interpretation of "big diagnostic data" and the plethora of patient-oriented, environmental and epidemiological clinical data. Moreover we have discovered a set of novel genome typing markers using genome wide association studies in order to introduce a robust alternative of classical typing method. The ultimate target for such approaches is to streamline and accelerate data management in favor of improved patient care.

## Introduction

Over many hundreds of millions of years every conceivable ecological niche on the planet earth has become inhabited by microbes such as bacteria, viruses and fungi. Indeed, microbes are highly adaptable to external selective forces and ubiquitous on Earth. Bacteriophages, which are viruses infecting bacteria, represent the most abundant life form in the biosphere. Many microbes are essential to human, animal and plant life. In contrast, many microbial species have been identified as a pathogen because they cause acute infectious diseases or trigger chronic diseases. Hantavirus pulmonary syndrome caused by Sin Nombre Hanta virus, viral encephalitis from Nipah virus, tuberculosis by *Mycobacterium tuberculosis,* cholera by *Vibrio cholera, Clostridium difficile* infection, infection by *Pseudomonas aeruginosa*, and the most recent pandemic called Coronavirus disease 2019 (COVID19) caused by *severe acute respiratory syndrome coronavirus* 2 (SARS-CoV-2) are just a few examples of severe infections that affect human health. Rapid detection and epidemiological characterization of pathogenic microbes is highly essential to control any disease outbreak situation or to specifically and precisely treat microbial infection. Inaccurate identification of microbes frustrates correct taxonomic classification and can misdirect proper treatment of infectious diseases caused by microorganisms (Franco-Duarte et al., 2019). However, in clinical microbiology, diagnostics needs to start with an accurate identification of pathogenic microbes. Continuous technological advancement provided a variety of methods from classical to the most recent technologies for strain detection and characterization (Franco-Duarte et al., 2019, Ferone et al., 2020). In this review we are primarily focused on bacterial typing and characterization.

Bioinformatics is an advanced and relatively recent discipline that exploits computational methodologies which integrate applied mathematics and statistics to the study of biological phenomena to help solve scientific problems. Specifically, bioinformatics applications are becoming an imperative part of research in the fields of microbiology and infectious diseases, mostly as a consequence of continuously increasing availability of huge amounts of nucleic acid sequencing data gathered via high-throughput sequencing (HTS) technologies, more widely known as next-generation sequencing (NGS) (Goodwin et al., 2016, Mardis, 2013). NGS performed directly on clinical specimens generates information on presence and species of (a) pathogen(s), its / their antimicrobial resistance and virulence profiles, while at the same time providing detailed evolutionary and typing information that

supports epidemiological investigations on the origin of microbial types (Ma et al., 2014, Mintzer et al., 2019, Mitchell and Simner, 2019). This integral data package is unprecedented: there is not a single in vitro diagnostic (IVD) technology that generates data so complete and which provides for optimized management of infections. Next to facilitating these so-called metagenomic sequencing tests, NGS also allows for the complete elucidation of genome sequences for organisms that were purified by selective cultivation before. The whole genome sequencing (WGS) data generated by NGS permits even more comprehensive epidemiological study of microbial pathogens in the sense that isolated genome sequencing may be more complete (e.g. close to 100% genome coverage) than sequences obtained by metagenomics. Moreover, NGS technologies are critically and irreversibly changing the way in which microbial genomic material is analyzed. The technologies will replace classical, targeted molecular methods for microbial detection and typing, by generating a complete genome from of raw reads after assembly into a small number of contigs (Carrico et al., 2018). However, clinical microbiology laboratory testing is still mostly reliant upon demonstrable growth of a viable infectious microorganism.

Rapid detection and accurate identification of microorganisms in clinical specimens to be tested in the diagnostic microbiology laboratory has made constant progress in the various areas including bacteriology, mycology, mycobacteriology, parasitology and virology. As such classical microbiological technologies have had and still have a remarkably strong position in IVD and involve various procedures including conventional culture methods, immunological protocols (eg; ELISA), molecular methods and the more recently introduced spectroscopic methods (e.g., matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF MS)) (Ferone et al., 2020). Recent molecular methods of strain identification include fluorescent in-situ hybridization, nucleic acid amplification methods (polymerase chain reaction (PCR), quantitative qPCR, and reverse transcriptase rt-PCR but also technologies such as LAMP or Q-beta polymerase-based assays) (Fenollar and Raoult, 2004) and DNA microarrays (Ehrenreich, 2006). For instance, a widely accepted microbial characterization method called ribotyping is based on PCR amplification and restriction analysis of the intergenic spacer region (ISR) between 16S and 23S ribosomal RNA. The quality of molecular diagnostics can be strongly affected by specimen collection and transportation as well as the available laboratory equipment which is usually high cost and in need of hands on precision. In the past two decades clinical microbiology has evolved in parallel with increasingly relevant bioinformatic qatools for more detailed data analysis, assembly, interpretation and display (Hogeweg, 2011). This has resulted in the accumulation

of huge amounts of sequence data and published studies that created a new field called genomic epidemiology of pathogenic microbes. For example, a total of 246,189 (complete and draft) bacterial and 6,615 viral genomes have been deposited in an online genome database called GOLD (Mukherjee et al., 2017). Unlike the classical microbiological typing techniques that only cover a limited number of phenotypic or genomic markers, genomic epidemiological analysis of NGS data allows one to obtain whole genome-scale insights of pathogens using variety of bioinformatics applications and tools (Maiden et al., 2013). The continuously increasing volume of NGS data stimulated the development of online databases (DB) related to microbial typing. This covered Multi Locus Sequence Typing (MLST) DBs, spa-typing DBs, genomic serotyping DBs as well as DBs integrating antimicrobial resistance (AMR) genes and virulence genes (Carriço et al., 2013). Furthermore, the application of WGS data analysis for infectious microbial agents is promising for optimizing treatment of infectious diseases via accurate strain typing in order to manage pathogen outbreaks (Editors, 2002, Boers et al., 2012).

Below we will briefly sketch the classical typing technologies still used in routine clinical microbiology, the new typing technologies that were recently introduced and the associated bioinformatics modules that facilitate the management of increasingly large amounts of diagnostic and clinical data, often in combination.

## MICROBIAL DETECTION AND CHARACTERIZATION

- **The conventional microbiology technologies**

  Conventional methods of bacterial detection are traditionally based on cultivation, including sample preparation, enrichment, dilution, plating, enumeration, and isolation of single species colonies for more detailed characterization (Nomura M, 1999, Gracias and McKillip, 2004, Ferone et al., 2020). Basic principles of conventional identification are based on both morphological (different colony size and color), biochemical, physiological and genetic characteristics of the microorganisms involved. Selective culture media can be used for enumeration of specific classes of pathogenic bacteria as present in complex mixtures (e.g. feces or nasopharyngeal swabs) and classical microbial taxonomic techniques are used to determine the identity of the cultured organisms (Lau et al., 2003). Their putative pathogenicity is further investigated by biochemical or serological tests (Gugliandolo et al., 2011, Váradi et al., 2017). Such phenotype-based methods are not only time

consuming, qualitative more than quantitative, cumbersome and less reliable but may also require specific reagents that are usually not universally available (Yeung et al., 2002, Zhao et al., 2014). However, despite these limitations and the availability of more advanced techniques for microbial identification, phenotype-based or culture-dependent conventional methods are still considered as gold standard technologies in the detection of major human microbial pathogens (Donelli et al., 2013). Tests based on the analytical profile index (API), where standard biochemical methods mentioned above are integrated into miniaturized reaction vessels and scored as "positive" or "negative", generate a profile that is characteristic for certain species. The automated versions of such a test (e.g. the Vitek solutions (bioMérieux, Craponne, France)) facilitate the more rapid and precise reading of test reactions and are a perfect example of an improved version of a relatively recent culture-dependent conventional method (Funke and Funke-Kissling, 2004, Sutton, 2007).

- **Nucleic acid amplification technologies (Molecular techniques):**

  During the last four decades a large number of molecular methods have been developed to overcome the limitations of conventional microbiological methods and to facilitate the rapid, accurate, sensitive and cost effective identification and enumeration of microorganisms, especially for non- or poorly-cultivable ones (Galluzzi et al., 2007, HÖFLING et al., 1997, Spratt, 2004).

  These methods are culture independent and based on nucleic acid hybridization and fragment-based amplification. Classical ribotyping is a classic example of such a hybridization method (Grimont and Grimont, 1986). It is based on the identification of a number of ribosomal gene loci and their position in the chromosome. Although it is highly reproducible and particularly applicable to fast-growing bacteria, it still has a relatively low discriminatory power (Blanc et al., 1994, Pfaller et al., 1996). However, due to its robustness and reproducibility, inter-laboratory analyses of profiles are possible and this can be used to generate profile databases to help automatize ribotyping (Arvik et al., 2005). Among fragment-based methods, restriction fragment length polymorphism (RFLP) methods were widely used and based on the restriction-enzymatic digestion of chromosomes into several hundreds of small fragments, which are then separated by horizontal gel electrophoresis into complex banding patterns (Owen, 1989). This method is highly reproducible and robust but the complex restriction patterns hinder the inter-

laboratory sharing of data. However, by coupling this method with other hybridization techniques, data can be exchanged between different laboratories and can be integrated into central databases as well (Van Embden et al., 1993). Another gel electrophoresis technique, Pulsed-Field Gel Electrophoresis (PFGE) is still a popular method to separate large genomic fragments (up to hundreds of kilobases in length) in agarose gel by periodic alternation of the angle of the electric field's direction. PFGE has a remarkable discriminatory power and reproducibility, and has therefore become a widely applicable gold standard method for comparative typing of almost all bacterial species (Barg and Goering, 1993, Seifert et al., 2005). PFGE generates complex restriction patterns which needs dedicated softwares and databases for PFGE data interpretation.

Diagnostic PCR became most popular over the past two decades. PCR methods range from relatively simple or classical DNA amplification-based approaches targeting conserved regions in pathogen genomes flanked by primers from which exponential DNA synthesis originates. Despite the fact that PCR exhibits an adjustable level of discrimination, flexibility, technical simplicity and broad availability, PCR 'fingerprinting' data, in general, are considered to be non-exchangeable among laboratories. Random Amplification of Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP) and Multi Locus variable tandem repeats amplification (MLVA) are all PCR-mediated methods for bacterial strain typing (Ferone et al., 2020, Franco-Duarte et al., 2019, Yang and Rothman, 2004). Despite of being able to do simultaneous and reliable strain identification and in some instance quantification, specifically with reverse transcriptase quantitative PCR (RT-qPCR) (Bruce et al., 2020, Corman et al., 2020), molecular methods show some drawbacks related to the risk of contamination due to which false positive results may occur as well they are quite time consuming. Slightly different and faster techniques such as RT loop mediated isothermal amplification (RT-LAMP) of DNA have been used for high throughput tests for SARS-CoV-2 diagnosis in the recent pandemic (Gray et al., 2020, Jiang and Shi, 2020, Notomi et al., 2000, Zhang, 2020). However RT-LAMP rapid testing was observed to be less sensitivity, false negative results were relatively frequent, and typing applications have been rarely described (Ben-Assa et al., 2020, Butler et al., 2020). Given this situation, microbiologists developed a novel, rapid and more sensitive alternative of the methods listed above namely the PCR-LAMP technique which is a hybrid method

combining PCR-based amplification with isothermal amplification (Varlamov et al., 2020). The success rate and the feasibility of molecular technologies mentioned above depends on several factors including sample type (single or mixed-species), accuracy of results generated, resources and cost factors, as well as the turn-around-times expected. It is essential to understand the basic principles of these molecular methods, their precision and handling, their required instrumentation as well as other limitations. The complex PCR fingerprints are generally highly reproducible within institutes where protocols are well-respected and have been widely used successfully for high-throughput molecular typing of large numbers of bacterial isolates. Of note, without computer-assisted analyses, strain relatedness cannot be defined with sufficient precision (Table 1).

- **Mass spectrometry for microbial species identification:**

More recently, mass spectroscopy (MS) came into the diagnostic picture due to its practical simplicity and high accuracy, sensitivity and capability to detect even single nucleotide difference within a mixed population of strains of the same species as compared to amplification based techniques (Manukumar and Umesha, 2017, Sauer and Kliem, 2010, Singhal et al., 2015). The basic principle of MS is based on the detection of an intrinsic physical property, the mass-to-charge ratio (m/z) of a bioanalyte such as nucleic acids, proteins, lipids, carbohydrates and peptides, which makes MS a unique method of microbe detection and characterization over conventional and molecular techniques (Sandrin and Demirev, 2018, Sauer and Kliem, 2010). Since many years MS has been using in combination with other ionization and separation techniques. Gas chromatography (GC) (Nichols et al., 1986, Senes et al., 2018), matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) (Jang and Kim, 2018, Rahi et al., 2016, Schauer et al., 2005), electromigration techniques/capillary electrophoresis (CE) (Desai and Armstrong, 2003, Lantz et al., 2007) and electrospray ionization (ESI) (Sampath et al., 2007, Smith et al., 1995) are actively used for microbe detection.

Bacterial strain identification using MALDI-TOF-MS is highly popular and rapid method among all other methods mentioned above. However, MALDI-TOF-MS is not only culture dependent but sometime also incapable of differentiating closely related strains such as for example different *Streptococcus* species whereas also

*Shigella* spp and *Escherichia coli* are usually indistinguishable (Ferone et al., 2020). Since MALDI-TOF-MS mostly generates unique spectra at the bacterial species level which need to be compared with well-characterized microorganisms from databases, MALDI-TOF-MS is highly dependent on next generation techniques of bioinformatics (Liébana-Martos, 2018, Oros et al., 2020). Mass spectra are calibrated to detect relevant peaks and then matched against a MS spectral fingerprints database of previous outbreak strains with the help of software packages (Table 1). The output is presented as dendograms in which isoproteomic strains are clustered closely together. Clustering is based on similarity of mass spectra from different sample sets, dates, and instruments. When such similarities are revealed between mass spectra, peaks of interest for further investigation can be highlighted (Christner et al., 2014, Dinkelacker et al., 2018, Oberle et al., 2016).

Surface enhanced Raman spectroscopy (SERS) is another rapid and highly sensitive method of bacterial identification allowing the monitoring of phenotypic changes of bacteria in response to different stress types such as the presence of antibiotics, heavy metals, toxic nanoparticles and starvation (Chang et al., 2019, Cui et al., 2019, Wang et al., 2016). Although the methodology holds great potential for high throughput detection and characterization, it still requires improvement to be recognized as a convenient microbial detection and typing technique. It requires standardization of technical parameters of spectral acquisition, consolidation of scattered in-house databases of spectral profiles, further development of innovative SERS substrates (Cui et al., 2019). However, all the spectroscopic fingerprints need bioinformatic interpretation to adequately define subtypes on the basis of the mass spectra recorded (Table 1).

***Table 1-1:*** *Bioinformatics playing significant role in result assessment of some molecular methods of microbiological strain typing . Some examples justifying the role are shown in the table.*

| **Classical microbial typing techniques** | **The use of Bioinformatic techniques /databases** | **Comment** | **References** |
|---|---|---|---|
| PFGE profiles | PulseNet (http://www.cdc.gov/pulsenet/) OR<br><br>SalmGene databases (http://www.hpa-bioinformatics.org.uk/bionumerics/salm_gene/) | Computer-assisted cluster analysis is inevitable for the comparison of large numbers of PFGE generated patterns | *(Seifert et al., 2005, Van Belkum et al., 1998)* |

| | | | |
|---|---|---|---|
| | (http://www.eurosurveillance.org/em/v10n10/1010-225.asp) | | |
| PCR generated fingerprints | GelJ tool embedded with PCR banding pattern database, GelComparII, and Phoretix 1D Pro | Cluster analysis of PCR band based fingerprints | *(Heras et al., 2015)* |
| MALDI-TOF MS / SERS | FlexControl software (Bruker Daltonics, Bremen, Germany), BioNumerics (Applied maths, Belgium) | Detailed analysis of the complex mass spectra profiles including thousands of peaks. | *(Oberle et al., 2016)* |
| Strain typing methods (MLST/ Spa typing/ cgMLST/ wgMLST) | MLST.net<br><br>SeqNet.org spa sequence repository (spaserver.ridom.de)<br><br>(http://pubmlst.org/)<br><br>(SeqSphere+ software v5.1.0, Ridom GmbH, Münster, Germany; BioNumerics, Applied Maths, Belgium) | MLST (based on seven locus) and cgMLST/ wgMLST schemes are developed using whole genome sequences of microbes | *(Jolley et al., 2004)* |

- **Next generation "omics" technologies**

All current methods for microbial detection and characterization have the potential to offer reliable results but also carry limitations. Steady evolution of sequencing technologies, from Sanger sequencing to contemporary NGS combined with several other omics technologies (e.g. transcriptomics, metabolomics and proteomics), will provide a better understanding of the physical composition of microbes. The clinical application of combined omics will also create a solid foundation to uncover new genomic information associated with health hazards. These new insights will generate breakthroughs in the understanding of bacterial genomics in terms of survival mechanisms, generation and frequency of mutations, virulence characteristics, increasing drug resistance and more generic features of microbial pathogenesis (Bostanci et al., 2019, Cocolin et al., 2018, den Besten et al., 2018, Dylus et al., 2020, Goldberg et al., 2015, Pulido et al., 2016, Van Goethem et al., 2019, Schneider and Orchard, 2011, Quainoo et al., 2017). Moreover, traditional experimental techniques in combination with omics tools strengthen the understanding of complex biological dynamics and microbial risk assessment (Karahalil, 2016). For instance, recent work by (Kuijpers et al., 2018) demonstrates variability in infectivity for different *Salmonella* strains obtained from in vitro gastro-

intestinal tract infection experiments. However, when such experiments were coupled with subsequent NGS-omics studies in Salmonella strains this helped establishing a true biological dose-response relationship (Haddad et al., 2018).

- **NGS technologies and metagenomics**

  Application of current culture-dependent and analytical microbiological diagnostics is not sufficient for outbreak and transmission investigations. NGS relies on high-throughput WGS that produces millions of sequence runs in one go. Given the fact that NGS is capable of extracting complete genome sequences, these data have multiple biological impacts on microbial strain typing, assessment of relatedness, detailed epidemiological characterization, identification of antimicrobial resistance (AMR) genes, virulence genes and surveillance of outbreaks of infections in hospitals or the community (Abdelbary et al., 2019, Couto and Rossen, 2021, Deurenberg, 2017, Dunne et al., 2012, Mellmann et al., 2017). In addition, NGS provides a revolutionary way to define new markers related to AMR and strain typing *(Goldberg et al., 2015, Goyal et al., 2020, Tshibangu-Kabamba et al., 2020, Van Goethem et al., 2019)*. The application of cutting edge NGS technology to study sequence data derived from a complex sample containing several microorganisms is commonly known as metagenomics (Couto and Rossen, 2021, Wooley et al., 2010). Unlike the single isolate WGS, metagenomics can be separated in two principally different applications. First, amplicon-based metagenomes were investigated via the 16S-23S rRNA encoding region as target. This approach was and is still being used for concurrent identification of several pathogens in samples to allow detection of all species present (Couto and Rossen, 2021, Sabat et al., 2017). Second, shotgun metagenomics (Couto and Rossen, 2021) allows for the definition of full genomic sequences for microbes present in a sample. All nucleic acid molecules present are to be sequenced and this will give a complete review of all genomes present. Analysis of the complex data thus obtained is done using different bioinformatics tools in order to generate phylogenetic relationship maps. Recently, the successful identification and typing of Dengue virus was done using an optimized version of shotgun metagenomics (Lizarazo et al., 2019). Furthermore, shotgun metagenomics is a highly popular method to characterize the gut microbiome or environmental microbiomes (Chiu and Miller, 2019, Gigliucci et al., 2018, Wooley et al., 2010, Zhao and Bajic, 2015).

## INNOVATIVE NGS-BASED BIOINFORMATIC APPROACHES

Spectacular advancement of biological and information processing science and its amalgamation with computer algorithms has generated a promising perspective on the future availability of innovative healthcare systems in several medical fields. Interactive data mining in research laboratories and hospital information systems has drawn a broader picture of clinical findings which is not only important to monitor clinical history of diseases/ infections but also to deal with emerging health risks and effective implementation of genomics medicines with an overall target of improving patient care (Saeb, 2018). Although molecular methods discussed above are the current gold standard in the field of infection diagnostics, interpretation of test results cannot always rely on phenotypic data and thus needs integration with bioinformatics approaches to overcome molecular diagnostic limitations such as time, cost, lack of reproducibility of the results and reliability. Interpretation of experimental diagnostic outcomes produced by molecular methods (Table1-2), consolidation of databases of different biological signatures, machine learning (ML) and NGS data analysis are being increasingly accepted in the field of diagnostics and infection control. Recently, ML approaches have been applied to NGS data in order to identify unknown pathogens as well as unknown markers associated with different phenotypes such as drug resistance, pathogenicity and strain types (Luz et al., 2020, Shamout et al., 2021, Zou et al., 2019).

## SEQUENCING PLATFORMS AND BASIC WORKFLOW OF NGS DATA ANALYSIS

Parallel processing of clinical samples via NGS methods produces complex genomic sequence data. Availability of these massive datasets dramatically shifts the paradigm of epidemiology from targeted experimental methods to the processing of sequencing reads in order to explore complete genome information of pathogens. In public health microbiology and the therapeutic sector, NGS helps clinicians to take accurate decisions directed toward epidemiology and the treatment of infectious diseases (Fricke and Rasko, 2014, Maljkovic Berry et al., 2020). A recent milestone of NGS technology was the generation of more than 1.2 million coronavirus genome sequences from 172 countries during the ongoing COVID-19 pandemic (Maxmen, 2021).

In 1993, Tuberculosis (TB) was declared a “global health emergency” by the World Health Organization (WHO). At that time molecular diagnostic tests simply relied on

phenotypic methods for early detection and identification of *Mycobacterium tuberculosis* and their resistance to first-line drugs (Richeldi, 2006). Classical diagnostic methods were then key tools in the containment of multi drug resistant (MDR) TB. Rapid development in the field of diagnostic microbiology has opened up for the application of NGS technologies to a new dimension molecular epidemiology of tuberculosis, in which the transmission chain of infection at higher resolution could be easily traced (Comas, 2017).

The NGS approach generally starts with genomic DNA extraction from test samples, cDNA library preparation which involves DNA fragmentation, ligation of adaptors, adaptor sequencing, and sample enrichment, sequencing using dedicated instruments and bioinformatics analysis (Figure 1-1) (Buermans and Den Dunnen, 2014, Grada and Weinbrecht, 2013).

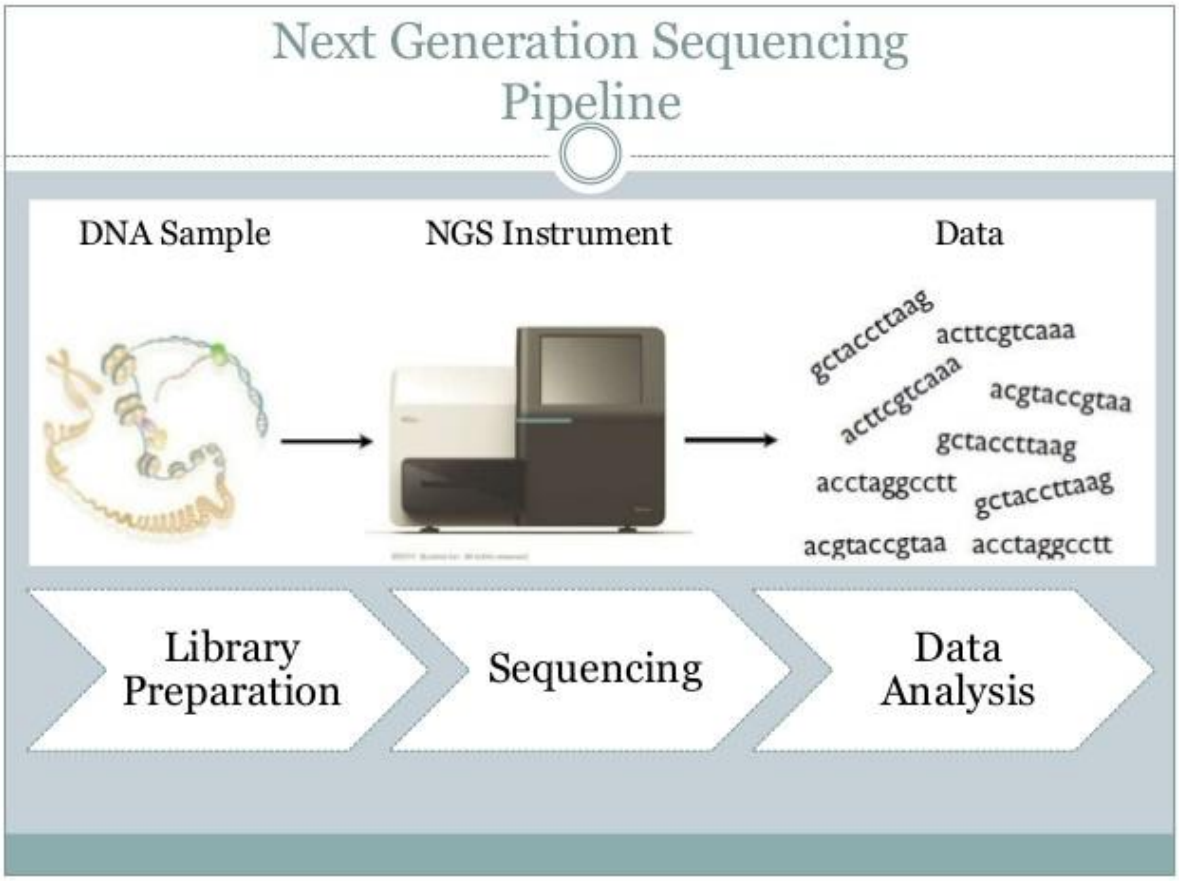



***Figure 1-1**: Next generation sequencing workflow scheme.*

Benchtop NGS platforms with different properties such as output quality, quantity and fragment/read length are currently available (Table 1-2) (Heather and Chain, 2016). MiSeq, NextSeq/ HiSeq and Ion Torrent are considered as suitable platforms for smaller targets (bacteria /virus) whereas for highly repetitive bacterial genomes, with modular

plasmid structures, HiSeq and Novaseq platforms are suggested (Faria et al., 2016, Gire et al., 2014, LaBreck et al., 2018, Salje et al., 2017, Stewart-Ibarra et al., 2018). However, the Illumina HiSeq family of NGS platforms is widely preferred for diagnosis and public health by large scale companies as well as laboratories (Chen et al., 2021, Liu et al., 2012, Rhodes et al., 2014). Recent innovation in nanopore sequencing from Oxford Nanopore Technologies (ONT) showed the potential to challenge other sequencing platforms. Due to its low cost, rapid turnaround time, and user-friendly bioinformatics pipelines, ONT sequencing becomes an attractive platform for clinical laboratories to adopt. However, this method still faces the problem of base-calling accuracy compared to other platforms (Petersen et al., 2019). In 2011, Pacific Biosciences introduced the first PacBio RS sequencing platform (using first generation chemistry, P1-C1) to the market. This machine uses single molecule real-time (SMRT) detection technology that achieves real-time sequencing of individual polymerase molecules (Eid et al., 2009, La et al., 2021, Teng et al., 2017) (Table 1-2). One of the most common strategies for maximizing efficiency of a sequencing technology is the multiplexing of samples; a unique index is appended to each sample, and multiple samples are pooled together for sequencing in the same run due to which sequencing platforms suffer from the index swapping issue. However, BGI sequencing services offers a unique DNA nanoball (DNB) technology Rolling circle replication (RCR) amplification that has rare background-level single index mis-assignment during DNB preparation and library construction (Cui et al., 2019). In the present times Illumina, PacBio, ONT and BGI are among the preferred technologies in the field of metagenomics. However, all the sequencing technologies have some pros and cons as well thus the selection should be based on the purpose of study.

***Table 1-2:*** *Some popular NGS platforms with their underlying technologies, advantages and disadvantages.*

| **Sequencing platform** | **Underlying Chemistry** | **Read length (base pair: bp)** | **Run time** | **Advantage** | **Disadvantage** |
|---|---|---|---|---|---|
| **Illumina (MiSeq/HiSeq)** | Sequencing by synthesis | 50-250bp | 1 to 10 day depending upon sequencer, read length and mode | High throughput/ less expensive | Short reads and long run time in default mode |
| **ThermoFisher scientific (Ion Proton/Ion** | Proton detection | 200bp | 2 hrs | Short run time, less expensive | Homopolymer errors |

| | | | | | |
|---|---|---|---|---|---|
| **Torrent)** | | | | | |
| **Roche 454** | Pyrosequencing | 700 bp | 24 hrs | Long read length, Fast | High homopolymer error rate high cost, low throughput |
| **SOLiD sequencing** | Ligation and two-base coding | 85-100 bp | 1 to 2 weeks | Low cost per base, accuracy | Short reads and slowest method |
| **Pacific Biosciences (PacBio)** | Sequencing by synthesis: SMRTbell replication | ≥500 bp | Up to 30 hrs | Long read length, Fast | High capital cost, variable accuracy |
| **ONT (Oxford Nanopore Technologies)** | Measures the changes in current as biological molecules pass through the nanopore | ≥500 bp | Up to 72 hrs | Long reads, low capital cost, fast | Low accuracy |
| **DNBSEQ from BGI** | DNA nanoball/ Rolling circle replication (RCR) | 50 to 150 bp | 24 to 30 hrs | Accurate, Fast, Flexible | Short reads |

## NGS DATA ANALYSIS

NGS data analysis starts with checking the quality of sequence reads. The Phred quality score (Q score) measures the probability of incorrect base calling which is defined as a logarithmic probability of base calling errors (Ewing and Green, 1998). For instance, generally Q30 score is considered as an ideal Q score for Illumina sequencing reads. In the next step adapters attached to the reads are removed by trimming. Trimmed high quality reads are then subjected to assembly (*de novo* or reference based). In recent years, key players in the field of microbial genome assembly are CLCbio workbench, GENEIOUS, SPAdes, DNASTAR by Lasergene and EvoCAT (Bankevich et al., 2012, Goyal et al., 2020, Segerman, 2020, Goyal et al., 2019, Souvorov et al., 2018). Illumina like platform generates accurate but short reads, which can lead to accurate but fragmented genome assemblies whereas PacBio and ONT like platforms generate long reads that can produce complete

genome assemblies, but more expensive and error-prone sequencing. Therefore hybrid assemblies provided by Unicycler can be a significant option which combines data from complementary sequencing technologies to generate more accurate assembly (Liu et al., 2020, Wick et al., 2017). The quality of the assemblies is evaluated by QUAST (Gurevich et al., 2013) and contigs below 200 bp in length are discarded. Additional genome annotation is performed using an annotated reference genome using different pipelines such as CLCbio workbench, BioNumerics and Prokka (Goyal et al., 2019, Seemann, 2014). Assembled and annotated genomes are further analyzed for epidemiological typing and characterization of the genomes.

**Genome typing**

Multi Locus Sequence Typing (MLST) is a commonly used classical approach for epidemiological typing of pathogens such as *Pseudomonas aeruginosa*. It accurately defines evolutionary descent and identifies distinct lineages but it lacks the necessary resolution for the characterization of outbreaks caused by closely related, contemporaneous bacterial isolates (Ashton et al., 2016a, Inns et al., 2015b). Several studies have evaluated the discriminatory power and concordance of different typing methods (Gateau et al., 2019a, Rumore et al., 2018a). However, high throughput WGS is rapidly becoming the most efficient solution for strain typing, both for surveillance and for (retrospective) outbreak investigations (Kan et al., 2018a). WGS facilitates whole genome MLST (wgMLST) which displays higher discrimination than conventional MLST which is usually based on variation in seven housekeeping genes. Most bacterial species have sufficient variation within house-keeping genes to provide many alleles per locus, allowing billions of distinct allelic profiles to be distinguished using only seven house-keeping genes. wgMLST reliably recognizes and quantifies the genetic links between epidemiologically related isolates within various bacterial species (Joensen et al., 2014a, Kovanen et al., 2014a). Additionally, genetic maps can be drawn by performing phylogenetic analysis using WGS of microbe's population.

**Genome Wide Association Studies: SNPs vs k-mers as identity markers**

Whole genome Single Nucleotide Polymorphisms (wgSNPs) based genotyping is one of the most advanced methods of exploiting conserved as well as variable regions on whole genome level in order to identify transmission dynamics and to provide useful insights into the sources and routes of infection during hospital outbreaks which will not necessarily be

concordant with core genome analysis solely (den Bakker et al., 2011, Halachev et al., 2014b, Taylor and Unakal, 2021).

On the contrary, availability of WGS data and antibiotic resistance profiles of clinical strains enables genome wide association studies (GWAS) into the discovery of new potential target for antibiotics (Jaillard et al., 2018b, Lees et al., 2018). Past studies reported the successful demonstration of bacterial GWAS to recover known AMR determinants as well as to formulate new hypotheses involving genetic variants not yet described in the antibiotic resistance literature (Jaillard et al., 2018b). De Bruijn graph GWAS (DBGWAS) allows the identification of short DNA fragments (signature sequences) associated to a given condition (such as MIC profile of all the strains against different antibiotics). De Bruijn graphs are built to connect overlapping k-mers, yielding a compact summary of all variations across a set of genomes (Jaillard et al., 2018b). Genetic variants (connected overlapping k-mers, called unitigs) are selected on the basis of their considerable association in desired phenotype and minimum q-value. Q-values are Benjamini-Hochberg transformed p-values for controlling the false positive results in the case of multiple testing (Benjamini and Hochberg, 1995, Jaillard et al., 2018b). Significant markers are further annotated using the online BLAST tool (www.blast.ncbi.nlm.nih.gov) (Goyal et al., 2020).

Methods other than DBGWAS such as PLINK (SNP based) and SEER (K-mer based) implemented in pyseer v 1.2.0 are also available to perform GWAS (Lees et al., 2018, Purcell et al., 2007, Saber and Shapiro, 2020). However, Scoary is another freely available python script to perform GWAS specifically for pan genomes thus termed as pan-GWAS which is based on the presence and absence of genes and their associated phenotypes in the dataset of pan genomes (Brynildsrud et al., 2016, Redfern et al., 2021).

**Big Data Challenges and Future Applications of Bioinformatics**

Health care-associated sectors are being flooded with NGS data and clinicians are trying to use genomic sequencing data for diagnosis and monitoring of infectious diseases. Analysis of big omics data needs advanced bioinformatic technologies and regularly updated and integrated databases containing information on molecular profiles, epidemiology and metadata of outbreak strains (Carriço et al., 2013, Saeb, 2018). Correctly storing and interpreting this huge amount of NGS data is a big concern for biologist among whom there is a strong need of cloud computing where data and high-power computing software are situated to be accessed virtually by users (Marx, 2013, Seth et al., 2019). In many

laboratories and companies cloud computing is a primary option and people rarely work on classical hardware components anymore. Undoubtedly, cloud computing will emerge to be a cost effective technique to process and accumulate the immense quantity of data with parallel processing tools and high protection storage through the internet (Wordsworth et al., 2018). However, sharing big voluminous genomic data and processing tools with outside collaborators using the cloud while maintaining internal confidentiality and the trust values of cloud service providers still remain major challenges to deal with (Raza and Luheshi, 2016, Seth et al., 2019). Some publically available big data cloud services include Google Cloud Platform (https://cloud.google.com/products), Amazon Cloud Services (http://aws.amazon.com), IBM Cloud Services www.ibm.com/cloud, which facilitate secure data access, migration, storage, retrieval, and computational processing. Big private organizations/ companies are also relying on their own cloud computing facilities to prevent the risk of data theft.

**Global Bioinformatics Market**

The globally exploding biological data resources have created platforms for several companies and service providers to manage and analyze complex data with the help of updated sophisticated computational techniques, algorithms and statistical methods. Such companies provide efficient means of storing, searching and retrieving the data for future infection and disease outbreak management. Data management systems apparently cover the entire bioinformatics data lifecycle including managing and monitoring the intake, integrity, and use of diverse bioinformatics data types. Development and implementation of policies, processes, and templates constituting an overarching data management plan supporting multiple platforms for large projects in collaboration with the customers are also taken care of by companies/ service providers. World's largest revenue impact advisory firm MarketsandMarkets™ stated that the global bioinformatics market is expected to account for USD 7,063.7 billion in 2018. It is expected to reach USD 13,901.5 billion by 2023, at a CAGR of 14.5% during the forecast period (https://www.marketsandmarkets.com/). The COVID-19 pandemic was a real-life example to show the potential of bioinformatic data analyses. Since the beginning of the COVID-19 pandemic, the major focus areas for every country have been to study and understand stopping progress of the virus. An important fact is that molecular biology has generated a vast amount of worldwide genomic data of the coronavirus. However, decoding the genome of novel coronavirus using bioinformatics tools

and algorithms was a crucial step in developing vaccines and for better understanding the infection mechanism used by the virus and how it functions in the human body (Goyal et al., 2021, Hufsky et al., 2021, Ishack and Lipner, 2021, Ray et al., 2021). According to the report published in July 2021 by MarketsandMarkets™, prominent players in the bioinformatics services market include Illumina (US), Thermo Fisher Scientific (US), Eurofins Scientific (Luxembourg), BGI Group (China), NeoGenomics (US), PerkinElmer (US), CD Genomics (US), Psomagen, Inc. (South Korea), QIAGEN (Germany), GENEWIZ (US), Source BioScience (UK), Microsynth (Switzerland), MedGenome (India), Fios Genomics (UK), and BaseClear (Netherlands), among others (https://www.marketsandmarkets.com/).

**Recent trends and futuristic approach for bioinformatics data analysis**

The accumulation of bioinformatics data and tools is coming to a turning point that requires a paradigm shift. Local data storage and analysis is reaching its limits, and the trend should be towards integrated and standardized cloud solutions that are fully automated to perform analysis using powerful computational algorithms. Data usually comes from multiple sources that need to be aggregated. It often arrives in batches and requires punctually large resources. Statistical analysis runs routinely on regularly updated data.

ML is a growing field in microbiology specially in data-intense discipline like genomics, it helps in diagnostics, classification, outcome prediction, antimicrobial risk management and to predict phenotypes from genotypes in infectious diseases *(Peiffer-Smadja et al., 2020, Zou et al., 2019).* Cloud solutions like AWS can help build powerful AI-driven pipelines that adapt to the flow of data and maximize its value.

**Conclusions**

Since the past two decades microbial epidemiology has been based on classical and advanced molecular techniques. However, interpretation of end results of high throughput molecular techniques are now dependent of bioinformatics applications. Molecular biology and bioinformatics techniques are complementary to each other. Everyday vast amounts of NGS data are being generated using molecular biology techniques and processed and analyzed by bioinformatics techniques or different purposes such as genome characterization, identification of antimicrobial drug resistance, phylogenetic analysis and disease outbreak surveillance and management etc. In this review we have succinctly described the inevitable need of better management of the steadily growing data repository along with molecular

biological advancements. Different microbiological techniques available for microbial detection and characterization were summarized in this review. On the other hand various roles played by NGS data analysis in the field of epidemiology and infection outbreak management was also shown in detail. Our focus was also to define the ways of big data management and future resources to analyze the giant repository of high-quality multi omics data mainly for the purpose of microbial disease epidemiology.

**Description of Thesis Activities**

I have been using bioinformatics approaches to better define the classical microbiological and molecular behavior and dynamics of strains of several bacterial species that are pathogenic to humans. I studied the genomic variations in resident strains of *Staphylococcus aureus* colonizing, sometimes for years, the nasal cavities of human volunteers. In addition we performed several genome wide association studies in which I tried to correlate known capacities of strains of *Clostridioides difficile* with genomic diversity among these strains. Finally, I studied collections of *Pseudomonas aeruginosa* strains and *SARS-CoV-2* strains using both state of the art bioinformatics tools for epidemiological analysis as well as "customer friendly" bioinformatics tools allowing lay persons to rapidly interpret WGS results for improvement of infection control. Details of these studies can be found in the subsequent chapters, whereafter I will try to synthesize my findings in a closing discussion covering all chapters as a whole.

**References**

ABDELBARY, M. H., SENN, L., GREUB, G., CHAILLOU, G., MOULIN, E. & BLANC, D. S. 2019. Whole-genome sequencing revealed independent emergence of vancomycin-resistant Enterococcus faecium causing sequential outbreaks over 3 years in a tertiary care hospital. *European Journal of Clinical Microbiology & Infectious Diseases,* 38**,** 1163-1170.

ARVIK, T., HENICK-KLING, T. & GAFNER, J. 2005. Automated genotyping of Saccharomyces cerevisiae using the RiboPrinter®. *International journal of food microbiology,* 104**,** 35-41.

ASHTON, P. M., NAIR, S., PETERS, T. M., BALE, J. A., POWELL, D. G., PAINSET, A., TEWOLDE, R., SCHAEFER, U., JENKINS, C. & DALLMAN, T. J. 2016a. Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ,* 4**,** e1752.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV,

M. A. & PEVZNER, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.

BARG, N. L. & GOERING, R. V. 1993. Molecular epidemiology of nosocomial infection: analysis of chromosomal restriction fragment patterns by pulsed-field gel electrophoresis. *Infection Control & Hospital Epidemiology,* 14**,** 595-600.

BEN-ASSA, N., NADDAF, R., GEFEN, T., CAPUCHA, T., HAJJO, H., MANDELBAUM, N., ELBAUM, L., ROGOV, P., DANIEL, K. A., KAPLAN, S., ROTEM, A., CHOWERS, M., SZWARCWORT-COHEN, M., PAUL, M. & GEVA-ZATORSKY, N. 2020. SARS-CoV-2 On-the-Spot Virus Detection Directly from Patients. *medRxiv***,** 2020.04.22.20072389.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological),* 57**,** 289-300.

BLANC, D., LUGEON, C., WENGER, A., SIEGRIST, H. & FRANCIOLI, P. 1994. Quantitative antibiogram typing using inhibition zone diameters compared with ribotyping for epidemiological typing of methicillin-resistant Staphylococcus aureus. *Journal of clinical microbiology,* 32**,** 2505-2509.

BOERS, S. A., VAN DER REIJDEN, W. A. & JANSEN, R. 2012. High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS One,* 7**,** e39630.

BOSTANCI, N., BAO, K., GREENWOOD, D., SILBEREISEN, A. & BELIBASAKIS, G. N. 2019. Periodontal disease: From the lenses of light microscopy to the specs of proteomics and next-generation sequencing. *Advances in Clinical Chemistry,* 93**,** 263-290.

BRUCE, E. A., HUANG, M. L., PERCHETTI, G. A., TIGHE, S., LAAGUIBY, P., HOFFMAN, J. J., GERRARD, D. L., NALLA, A. K., WEI, Y., GRENINGER, A. L., DIEHL, S. A., SHIRLEY, D. J., LEONARD, D. G. B., HUSTON, C. D., KIRKPATRICK, B. D., DRAGON, J. A., CROTHERS, J. W., JEROME, K. R. & BOTTEN, J. W. 2020. Direct RT-qPCR detection of SARS-CoV-2 RNA from patient nasopharyngeal swabs without an RNA extraction step. *PLoS Biol,* 18**,** e3000896.

BRYNILDSRUD, O., BOHLIN, J., SCHEFFER, L. & ELDHOLM, V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol,* 17**,** 238.

BUERMANS, H. & DEN DUNNEN, J. 2014. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,* 1842**,** 1932-1941.

BUTLER, D. J., MOZSARY, C., MEYDAN, C., DANKO, D., FOOX, J., ROSIENE, J., SHAIBER, A., AFSHINNEKOO, E., MACKAY, M., SEDLAZECK, F. J., IVANOV, N. A., SIERRA, M., POHLE, D., ZIETZ, M., GISLADOTTIR, U., RAMLALL, V., WESTOVER, C. D., RYON, K., YOUNG, B., BHATTACHARYA, C., RUGGIERO, P., LANGHORST, B. W., TANNER, N., GAWRYS, J., MELESHKO, D., XU, D., STEEL, P. A. D., SHEMESH, A. J., XIANG, J., THIERRY-MIEG, J., THIERRY-MIEG, D., SCHWARTZ, R. E., IFTNER, A., BEZDAN, D., SIPLEY, J., CONG, L., CRANEY, A., VELU, P., MELNICK, A. M., HAJIRASOULIHA, I., HORNER, S. M., IFTNER, T., SALVATORE, M., LODA, M., WESTBLADE, L. F., CUSHING, M., LEVY, S., WU, S., TATONETTI, N., IMIELINSKI, M., RENNERT, H. & MASON, C. E. 2020. Shotgun Transcriptome and Isothermal Profiling of SARS-CoV-2 Infection Reveals Unique Host Responses, Viral Diversification, and Drug Interactions. *bioRxiv*.

CARRIÇO, J., SABAT, A., FRIEDRICH, A. & RAMIREZ, M. 2013. Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. *Eurosurveillance,* 18**,** 20382.

CARRICO, J. A., ROSSI, M., MORAN-GILAD, J., VAN DOMSELAAR, G. & RAMIREZ, M. 2018. A primer on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect,* 24**,** 342-349.

CHANG, K.-W., CHENG, H.-W., SHIUE, J., WANG, J.-K., WANG, Y.-L. & HUANG, N.-T. 2019. Antibiotic susceptibility test with surface-enhanced Raman scattering in a microfluidic system. *Analytical chemistry,* 91**,** 10988-10995.

CHEN, K.-H., LONGLEY, R., BONITO, G. & LIAO, H.-L. 2021. A Two-Step PCR Protocol Enabling Flexible Primer Choice and High Sequencing Yield for Illumina MiSeq Meta-Barcoding. *Agronomy,* 11.

CHIU, C. Y. & MILLER, S. A. 2019. Clinical metagenomics. *Nat Rev Genet,* 20**,** 341-355.

CHRISTNER, M., TRUSCH, M., ROHDE, H., KWIATKOWSKI, M., SCHLUTER, H., WOLTERS, M., AEPFELBACHER, M. & HENTSCHKE, M. 2014. Rapid MALDI-TOF mass spectrometry strain typing during a large outbreak of Shiga-Toxigenic Escherichia coli. *PLoS One,* 9**,** e101924.

COCOLIN, L., MATARAGAS, M., BOURDICHON, F., DOULGERAKI, A., PILET, M.-F., JAGADEESAN, B., RANTSIOU, K. & PHISTER, T. 2018. Next generation microbiological risk assessment meta-omics: the next need for integration. *International Journal of Food Microbiology,* 287**,** 10-17.

COMAS, I. 2017. Genomic epidemiology of tuberculosis. *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control***,** 79-93.

CORMAN, V. M., LANDT, O., KAISER, M., MOLENKAMP, R., MEIJER, A., CHU, D. K., BLEICKER, T., BRUNINK, S., SCHNEIDER, J., SCHMIDT, M. L., MULDERS, D. G., HAAGMANS, B. L., VAN DER VEER, B., VAN DEN BRINK, S., WIJSMAN, L., GODERSKI, G., ROMETTE, J. L., ELLIS, J., ZAMBON, M., PEIRIS, M., GOOSSENS, H., REUSKEN, C., KOOPMANS, M. P. & DROSTEN, C. 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill,* 25.

COUTO, N. & ROSSEN, J. W. 2021. Overview of Microbial NGS for Clinical and Public Health Microbiology. *Application and Integration of Omics-powered Diagnostics in Clinical and Public Health Microbiology.* Springer.

CUI, L., ZHANG, D., YANG, K., ZHANG, X. & ZHU, Y. G. 2019. Perspective on Surface-Enhanced Raman Spectroscopic Investigation of Microbial World. *Anal Chem,* 91**,** 15345-15354.

DEN BAKKER, H. C., MORENO SWITT, A. I., CUMMINGS, C. A., HOELZER, K., DEGORICIJA, L., RODRIGUEZ-RIVERA, L. D., WRIGHT, E. M., FANG, R., DAVIS, M. & ROOT, T. 2011. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common Salmonella enterica subsp. enterica serovar Montevideo pulsed-field gel electrophoresis type. *Applied and environmental microbiology,* 77**,** 8648-8655.

DEN BESTEN, H. M., AMÉZQUITA, A., BOVER-CID, S., DAGNAS, S., ELLOUZE, M., GUILLOU, S., NYCHAS, G., O'MAHONY, C., PÉREZ-RODRIGUEZ, F. & MEMBRÉ, J.-M. 2018. Next generation of microbiological risk assessment: Potential of omics data for exposure assessment. *International journal of food microbiology,* 287**,** 18-27.

DESAI, M. J. & ARMSTRONG, D. W. 2003. Separation, identification, and characterization of microorganisms by capillary electrophoresis. *Microbiol Mol Biol Rev,* 67**,** 38-51, table of contents.

DEURENBERG, R. H. B. E. C. M. A. C. N. F. M. G.-C. S. K.-S. A. M. D. R. E. 2017. Application of next generation sequencing in clinical microbiology and infection prevention, . *Journal of Biotechnology,,* 243**,** 9.

DINKELACKER, A. G., VOGT, S., OBERHETTINGER, P., MAUDER, N., RAU, J., KOSTRZEWA, M., ROSSEN, J. W. A., AUTENRIETH, I. B., 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. *J Clin Microbiol,* 56.

DONELLI, G., VUOTTO, C. & MASTROMARINO, P. 2013. Phenotyping and genotyping are both essential to identify and classify a probiotic microorganism. *Microb Ecol Health Dis,* 24.

DUNNE, W., WESTBLADE, L. & FORD, B. 2012. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *European journal of clinical microbiology & infectious diseases,* 31**,** 1719-1726.

DYLUS, D., PILLONEL, T., OPOTA, O., WÜTHRICH, D., SETH-SMITH, H., EGLI, A., LEO, S., LAZAREVIC, V., SCHRENZEL, J. & LAURENT, S. 2020. NGS-based S. aureus typing and outbreak analysis in clinical microbiology laboratories: lessons learned from a Swiss-wide proficiency test. *Frontiers in microbiology***,** 2822.

EDITORS, T. 2002. Molecular typing of micro-organisms: at the centre of diagnostics, genomics and pathogenesis of infectious diseases? *Journal of Medical Microbiology,* 51**,** 7-10.

EHRENREICH, A. 2006. DNA microarray technology for the microbiologist: an overview. *Applied microbiology and biotechnology,* 73**,** 255-273.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P. & BETTMAN, B. 2009. Real-time DNA sequencing from single polymerase molecules. *Science,* 323**,** 133-138.

EWING, B. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research,* 8**,** 186-194.

FARIA, N. R., AZEVEDO, R., KRAEMER, M. U. G., SOUZA, R., CUNHA, M. S., HILL, S. C., THEZE, J., BONSALL, M. B., BOWDEN, T. A., RISSANEN, I., ROCCO, I. M., NOGUEIRA, J. S., MAEDA, A. Y., VASAMI, F., MACEDO, F. L. L., SUZUKI, A., RODRIGUES, S. G., CRUZ, A. C. R., NUNES, B. T., MEDEIROS, D. B. A., RODRIGUES, D. S. G., QUEIROZ, A. L. N., DA SILVA, E. V. P., HENRIQUES, D. F., DA ROSA, E. S. T., DE OLIVEIRA, C. S., MARTINS, L. C., VASCONCELOS, H. B., CASSEB, L. M. N., SIMITH, D. B., MESSINA, J. P., ABADE, L., LOURENCO, J., ALCANTARA, L. C. J., DE LIMA, M. M., GIOVANETTI, M., HAY, S. I., DE OLIVEIRA, R. S., LEMOS, P. D. S., DE OLIVEIRA, L. F., DE LIMA, C. P. S., DA SILVA, S. P., DE VASCONCELOS, J. M., FRANCO, L., CARDOSO, J. F., VIANEZ-JUNIOR, J., MIR, D., BELLO, G., DELATORRE, E., KHAN, K., CREATORE, M., COELHO, G. E., DE OLIVEIRA, W. K., TESH, R., PYBUS, O. G., NUNES, M. R. T. & VASCONCELOS, P. F. C. 2016. Zika virus in the Americas: Early epidemiological and genetic findings. *Science,* 352**,** 345-349.

FENOLLAR, F. & RAOULT, D. 2004. Molecular genetic methods for the diagnosis of fastidious microorganisms. *Apmis,* 112**,** 785-807.

FERONE, M., GOWEN, A., FANNING, S. & SCANNELL, A. G. 2020. Microbial detection and identification methods: Bench top assays to omics approaches. *Comprehensive Reviews in Food Science and Food Safety,* 19**,** 3106-3129.

FRANCO-DUARTE, R., ČERNÁKOVÁ, L., KADAM, S., S KAUSHIK, K., SALEHI, B., BEVILACQUA, A., CORBO, M. R., ANTOLAK, H., DYBKA-STĘPIEŃ, K. & LESZCZEWICZ, M. 2019. Advances in chemical and biological methods to identify microorganisms—from past to present. *Microorganisms,* 7**,** 130.

FRICKE, W. F. & RASKO, D. A. 2014. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics,* 15**,** 49-55.

FUNKE, G. & FUNKE-KISSLING, P. 2004. Evaluation of the new VITEK 2 card for identification of clinically relevant gram-negative rods. *J Clin Microbiol,* 42**,** 4067-71.

GALLUZZI, L., MAGNANI, M., SAUNDERS, N., HARMS, C. & BRUCE, I. J. 2007. Current molecular techniques for the detection of microbial pathogens. *Science Progress,* 90**,** 29-50.

GATEAU, C., DEBOSCKER, S., COUTURIER, J., VOGEL, T., SCHMITT, E., MULLER, J., MÉNARD, C., TURCAN, B., ZAIDI, R. S. & YOUSSOUF, A. 2019a. Local outbreak of Clostridioides difficile PCR-Ribotype 018 investigated by multi locus variable number tandem repeat analysis, whole genome multi locus sequence typing and core genome single nucleotide polymorphism typing. *Anaerobe,* 60**,** 102087.

GIGLIUCCI, F., VON MEIJENFELDT, F. A. B., KNIJN, A., MICHELACCI, V., SCAVIA, G., MINELLI, F., DUTILH, B. E., AHMAD, H. M., RAANGS, G. C., FRIEDRICH, A. W., ROSSEN, J. W. A. & MORABITO, S. 2018. Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients. *Front Cell Infect Microbiol,* 8**,** 25.

GIRE, S. K., GOBA, A., ANDERSEN, K. G., SEALFON, R. S., PARK, D. J., KANNEH, L., JALLOH, S., MOMOH, M., FULLAH, M. & DUDAS, G. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *science,* 345**,** 1369-1372.

GOLDBERG, B., SICHTIG, H., GEYER, C., LEDEBOER, N. & WEINSTOCK, G. M. 2015. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio,* 6**,** e01888-15.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17**,** 333-351.

GOYAL, M., DE BRUYNE, K., VAN BELKUM, A. & WEST, B. 2021. Different SARS-CoV-2 haplotypes associate with geographic origin and case fatality rates of COVID-19 patients. *Infection, Genetics and Evolution,* 90**,** 104730.

GOYAL, M., HAUBEN, L., POUSEELE, H., JAILLARD, M., DE BRUYNE, K., VAN BELKUM, A. & GOERING, R. 2020. Retrospective Definition of Clostridioides difficile PCR Ribotypes on the Basis of Whole Genome Polymorphisms: A Proof of Principle Study. *Diagnostics,* 10**,** 1078.

GOYAL, M., JAVERLIAT, F., PALMIERI, M., MIRANDE, C., VAN WAMEL, W., TAVAKOL, M., VERKAIK, N. J. & VAN BELKUM, A. 2019. Genomic evolution of Staphylococcus aureus during artificial and natural colonization of the human nose. *Frontiers in microbiology***,** 1525.

GRACIAS, K. S. & MCKILLIP, J. L. 2004. A review of conventional detection and enumeration methods for pathogenic bacteria in food. *Canadian journal of microbiology,* 50**,** 883-890.

GRADA, A. & WEINBRECHT, K. 2013. Next-generation sequencing: methodology and application. *J Invest Dermatol,* 133**,** e11.

GRAY, A. N., REN, G., ZHANG, Y., TANNER, N. & NICHOLS, N. 2020. Facilitating detection of SARS-CoV-2 directly from patient samples: precursor studies with RT-

qPCR and colorimetric RT-LAMP reagents. *Ipswich, MA: New England Biolabs. Application Note*.

GRIMONT, F. & GRIMONT, P. Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. Annales de l'Institut Pasteur/Microbiologie, 1986. Elsevier, 165-175.

GUGLIANDOLO, C., LENTINI, V., SPANÒ, A. & MAUGERI, T. 2011. Conventional and molecular methods to detect bacterial pathogens in mussels. *Letters in applied microbiology,* 52**,** 15-21.

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-5.

HADDAD, N., JOHNSON, N., KATHARIOU, S., MÉTRIS, A., PHISTER, T., PIELAAT, A., TASSOU, C., WELLS-BENNIK, M. H. & ZWIETERING, M. H. 2018. Next generation microbiological risk assessment—Potential of omics data for hazard characterisation. *International journal of food microbiology,* 287**,** 28-39.

HALACHEV, M. R., CHAN, J. Z., CONSTANTINIDOU, C. I., CUMLEY, N., BRADLEY, C., SMITH-BANKS, M., OPPENHEIM, B. & PALLEN, M. J. 2014b. Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant Acinetobacter baumanniiin Birmingham, England. *Genome medicine,* 6**,** 1-13.

HEATHER, J. M. & CHAIN, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107**,** 1-8.

HERAS, J., DOMINGUEZ, C., MATA, E., PASCUAL, V., LOZANO, C., TORRES, C. & ZARAZAGA, M. 2015. GelJ--a tool for analyzing DNA fingerprint gel images. *BMC Bioinformatics,* 16**,** 270.

HÖFLING, J. F., ROSA, E. A., BAPTISTA, M. J. & SPOLIDORIO, D. M. 1997. New strategies on molecular biology applied to microbial systematics. *Revista do Instituto de Medicina Tropical de São Paulo,* 39**,** 345-352.

HOGEWEG, P. 2011. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol,* 7**,** e1002021.

HUFSKY, F., LAMKIEWICZ, K., ALMEIDA, A., AOUACHERIA, A., ARIGHI, C., BATEMAN, A., BAUMBACH, J., BEERENWINKEL, N., BRANDT, C., CACCIABUE, M., CHUGURANSKY, S., DRECHSEL, O., FINN, R. D., FRITZ, A., FUCHS, S., HATTAB, G., HAUSCHILD, A. C., HEIDER, D., HOFFMANN, M., HOLZER, M., HOOPS, S., KADERALI, L., KALVARI, I., VON KLEIST, M., KMIECINSKI, R., KUHNERT, D., LASSO, G., LIBIN, P., LIST, M., LOCHEL, H. F., MARTIN, M. J., MARTIN, R., MATSCHINSKE, J., MCHARDY, A. C., MENDES, P., MISTRY, J., NAVRATIL, V., NAWROCKI, E. P., O'TOOLE, A. N., ONTIVEROS-PALACIOS, N., PETROV, A. I., RANGEL-PINEROS, G., REDASCHI, N., REIMERING, S., REINERT, K., REYES, A., RICHARDSON, L., ROBERTSON, D. L., SADEGH, S., SINGER, J. B., THEYS, K., UPTON, C., WELZEL, M., WILLIAMS, L. & MARZ, M. 2021. Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. *Brief Bioinform,* 22**,** 642-663.

INNS, T., LANE, C., PETERS, T., DALLMAN, T., CHATT, C., MCFARLAND, N., CROOK, P., BISHOP, T., EDGE, J. & HAWKER, J. 2015b. A multi-country Salmonella Enteritidis phage type 14b outbreak associated with eggs from a German producer:'near real-time'application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Eurosurveillance,* 20**,** 21098.

ISHACK, S. & LIPNER, S. R. 2021. Bioinformatics and immunoinformatics to support COVID-19 vaccine development. *J Med Virol,* 93**,** 5209-5211.

JAILLARD, M., LIMA, L., TOURNOUD, M., MAHÉ, P., VAN BELKUM, A., LACROIX, V. & JACOB, L. 2018b. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS genetics,* 14**,** e1007758.

JANG, K. S. & KIM, Y. H. 2018. Rapid and robust MALDI-TOF MS techniques for microbial identification: a brief overview of their diverse applications. *J Microbiol,* 56**,** 209-216.

JIANG, S. & SHI, Z. L. 2020. The First Disease X is Caused by a Highly Transmissible Acute Respiratory Syndrome Coronavirus. *Virol Sin,* 35**,** 263-265.

JOENSEN, K. G., SCHEUTZ, F., LUND, O., HASMAN, H., KAAS, R. S., NIELSEN, E. M. & AARESTRUP, F. M. 2014a. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of clinical microbiology,* 52**,** 1501-1510.

JOLLEY, K. A., CHAN, M. S. & MAIDEN, M. C. 2004. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics,* 5**,** 86.

KAN, B., ZHOU, H., DU, P., ZHANG, W., LU, X., QIN, T. & XU, J. 2018a. Transforming bacterial disease surveillance and investigation using whole-genome sequence to probe the trace. *Frontiers of Medicine,* 12**,** 23-33.

KARAHALIL, B. 2016. Overview of systems biology and omics technologies. *Current medicinal chemistry,* 23**,** 4221-4230.

KOVANEN, S. M., KIVISTÖ, R. I., ROSSI, M., SCHOTT, T., KÄRKKÄINEN, U.-M., TUUMINEN, T., UKSILA, J., RAUTELIN, H. & HÄNNINEN, M.-L. 2014a. Multilocus sequence typing (MLST) and whole-genome MLST of Campylobacter jejuni isolates from human infections in three districts during a seasonal peak in Finland. *Journal of Clinical Microbiology,* 52**,** 4147-4154.

KUIJPERS, A. F. A., BONACIC MARINOVIC, A. A., WIJNANDS, L. M., DELFGOU-VAN ASCH, E. H. M., VAN HOEK, A., FRANZ, E. & PIELAAT, A. 2018. Phenotypic Prediction: Linking in vitro Virulence to the Genomics of 59 Salmonella enterica Strains. *Front Microbiol,* 9**,** 3182.

LA, T.-M., KIM, J.-H., KIM, T., LEE, H.-J., LEE, Y., SHIN, H., SONG, Y., AHN, G., HUR, W., LEE, J.-B., PARK, S.-Y., CHOI, I.-S. & LEE, S.-W. 2021. The optimal standard protocols for whole-genome sequencing of antibiotic-resistant pathogenic bacteria using third-generation sequencing platforms. *Molecular & Cellular Toxicology,* 17**,** 493-501.

LABRECK, P. T., RICE, G. K., PASKEY, A. C., ELASSAL, E. M., CER, R. Z., LAW, N. N., SCHLETT, C. D., BENNETT, J. W., MILLAR, E. V., ELLIS, M. W., HAMILTON, T., BISHOP-LILLY, K. A. & MERRELL, D. S. 2018. Conjugative Transfer of a Novel Staphylococcal Plasmid Encoding the Biocide Resistance Gene, qacA. *Front Microbiol,* 9**,** 2664.

LANTZ, A. W., BAO, Y. & ARMSTRONG, D. W. 2007. Single-cell detection: test of microbial contamination using capillary electrophoresis. *Analytical chemistry,* 79**,** 1720-1724.

LAU, S. K., WOO, P. C., HUI, W.-T., LI, M. W., TENG, J. L., QUE, T.-L., LUK, W.-K., LAI, R. W., YUNG, R. W. & YUEN, K.-Y. 2003. Use of cefoperazone MacConkey agar for selective isolation of Laribacter hongkongensis. *Journal of clinical microbiology,* 41**,** 4839-4841.

LEES, J. A., GALARDINI, M., BENTLEY, S. D., WEISER, J. N. & CORANDER, J. 2018. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics,* 34**,** 4310-4312.

LIÉBANA-MARTOS, C. 2018. Indications, interpretation of results, advantages, disadvantages, and limitations of MALDI-TOF. *The Use of Mass Spectrometry Technology (MALDI-TOF) in Clinical Microbiology*. Elsevier.

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol,* 2012**,** 251364.

LIU, L., WANG, Y., CHE, Y., CHEN, Y., XIA, Y., LUO, R., CHENG, S. H., ZHENG, C. & ZHANG, T. 2020. High-quality bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly method. *Microbiome,* 8**,** 155.

LIZARAZO, E., COUTO, N., VINCENTI-GONZALEZ, M., RAANGS, E. C., VELASCO, Z., BETHENCOURT, S., JAENISCH, T., FRIEDRICH, A. W., TAMI, A. & ROSSEN, J. W. 2019. Applied shotgun metagenomics approach for the genetic characterization of dengue viruses. *J Biotechnol,* 306S**,** 100009.

LUZ, C. F., VOLLMER, M., DECRUYENAERE, J., NIJSTEN, M. W., GLASNER, C. & SINHA, B. 2020. Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. *Clinical Microbiology and Infection,* 26**,** 1291-1299.

MA, J., PRINCE, A. & AAGAARD, K. M. 2014. Use of whole genome shotgun metagenomics: a practical guide for the microbiome-minded physician scientist. *Semin Reprod Med,* 32**,** 5-13.

MAIDEN, M. C., JANSEN VAN RENSBURG, M. J., BRAY, J. E., EARLE, S. G., FORD, S. A., JOLLEY, K. A. & MCCARTHY, N. D. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol,* 11**,** 728-36.

MALJKOVIC BERRY, I., MELENDREZ, M. C., BISHOP-LILLY, K. A., RUTVISUTTINUNT, W., POLLETT, S., TALUNDZIC, E., MORTON, L. & JARMAN, R. G. 2020. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. *J Infect Dis,* 221**,** S292-S307.

MANUKUMAR, H. & UMESHA, S. 2017. MALDI-TOF-MS based identification and molecular characterization of food associated methicillin-resistant Staphylococcus aureus. *Scientific reports,* 7**,** 1-16.

MARDIS, E. R. 2013. Next-generation sequencing platforms. *Annual review of analytical chemistry,* 6**,** 287-303.

MARX, V. 2013. The big challenges of big data. *Nature,* 498**,** 255-260.

MAXMEN, A. 2021. POPULAR GENOME SITE HITS ONE MILLION CORONAVIRUS SEQUENCES. *Nature,* 593**,** 21.

MELLMANN, A., ANDERSEN, P. S., BLETZ, S., FRIEDRICH, A. W., KOHL, T. A., LILJE, B., NIEMANN, S., PRIOR, K., ROSSEN, J. W. & HARMSEN, D. 2017. High Interlaboratory Reproducibility and Accuracy of Next-Generation-Sequencing-Based Bacterial Genotyping in a Ring Trial. *J Clin Microbiol,* 55**,** 908-913.

MINTZER, V., MORAN-GILAD, J. & SIMON-TUVAL, T. 2019. Operational models and criteria for incorporating microbial whole genome sequencing in hospital microbiology - A systematic literature review. *Clin Microbiol Infect,* 25**,** 1086-1095.

MITCHELL, S. L. & SIMNER, P. J. 2019. Next-generation sequencing in clinical microbiology: are we there yet? *Clinics in laboratory medicine,* 39**,** 405-418.

MUKHERJEE, S., STAMATIS, D., BERTSCH, J., OVCHINNIKOVA, G., VEREZEMSKA, O., ISBANDI, M., THOMAS, A. D., ALI, R., SHARMA, K., KYRPIDES, N. C. & REDDY, T. B. 2017. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res,* 45**,** D446-D456.

NICHOLS, P. D., GUCKERT, J. B. & WHITE, D. C. 1986. Determination of monosaturated fatty acid double-bond position and geometry for microbial monocultures and complex consortia by capillary GC-MS of their dimethyl disulphide adducts. *Journal of Microbiological Methods,* 5**,** 49-55.

NOMURA M, K. H., SOMEYA Y, SUZUKI I. 1999. Novel characteristic for distinguishing Lactococcus lactis subsp. lactis from subsp. cremoris. *Int J Syst Bacteriol. ,* 49.

NOTOMI, T., OKAYAMA, H., MASUBUCHI, H., YONEKAWA, T., WATANABE, K., AMINO, N. & HASE, T. 2000. Loop-mediated isothermal amplification of DNA. *Nucleic acids research,* 28**,** e63-e63.

OBERLE, M., WOHLWEND, N., JONAS, D., MAURER, F. P., JOST, G., TSCHUDIN-SUTTER, S., VRANCKX, K. & EGLI, A. 2016. The Technical and Biological Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Based Typing: Employment of Bioinformatics in a Multicenter Study. *PLoS One,* 11**,** e0164260.

OROS, D., CEPRNJA, M., ZUCKO, J., CINDRIC, M., HOZIC, A., SKRLIN, J., BARISIC, K., MELVAN, E., UROIC, K., KOS, B. & STARCEVIC, A. 2020. Identification of pathogens from native urine samples by MALDI-TOF/TOF tandem mass spectrometry. *Clin Proteomics,* 17**,** 25.

OWEN, R. 1989. Chromosomal DNA fingerprinting—a new method of species and strain identification applicable to microbial pathogens. *Journal of Medical Microbiology,* 30**,** 89-99.

PEIFFER-SMADJA, N., RAWSON, T. M., AHMAD, R., BUCHARD, A., GEORGIOU, P., LESCURE, F. X., BIRGAND, G. & HOLMES, A. H. 2020. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect,* 26**,** 584-595.

PETERSEN, L. M., MARTIN, I. W., MOSCHETTI, W. E., KERSHAW, C. M. & TSONGALIS, G. J. 2019. Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing. *Journal of clinical microbiology,* 58**,** e01315-19.

PFALLER, M., WENDT, C., HOLLIS, R., WENZEL, R., FRITSCHEL, S., NEUBAUER, J. & HERWALDT, L. 1996. Comparative evaluation of an automated ribotyping system versus pulsed-field gel electrophoresis for epidemiological typing of clinical isolates of Escherichia coli and Pseudomonas aeruginosa from patients with recurrent gram-negative bacteremia. *Diagnostic microbiology and infectious disease,* 25**,** 1-8.

PULIDO, M. R., GARCÍA-QUINTANILLA, M., GIL-MARQUÉS, M. L. & MCCONNELL, M. J. 2016. Identifying targets for antibiotic development using omics technologies. *Drug discovery today,* 21**,** 465-472.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

QUAINOO, S., COOLEN, J. P., VAN HIJUM, S. A., HUYNEN, M. A., MELCHERS, W. J., VAN SCHAIK, W. & WERTHEIM, H. F. 2017. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clinical microbiology reviews,* 30**,** 1015-1063.

RAHI, P., PRAKASH, O. & SHOUCHE, Y. S. 2016. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass-Spectrometry (MALDI-TOF MS) Based Microbial Identifications: Challenges and Scopes for Microbial Ecologists. *Front Microbiol,* 7**,** 1359.

RAY, M., SABLE, M. N., SARKAR, S. & HALLUR, V. 2021. Essential interpretations of bioinformatics in COVID-19 pandemic. *Meta Gene,* 27**,** 100844.

RAZA, S. & LUHESHI, L. 2016. Big data or bust: realizing the microbial genomics revolution. *Microbial Genomics,* 2.

REDFERN, J., WALLACE, J., VAN BELKUM, A., JAILLARD, M., WHITTARD, E., RAGUPATHY, R., VERRAN, J., KELLY, P. & ENRIGHT, M. C. 2021. Biofilm associated genotypes of multiple antibiotic resistant Pseudomonas aeruginosa. . *BMC Genomics.,* 22**,** 26.

RHODES, J., BEALE, M. A. & FISHER, M. C. 2014. Illuminating choices for library prep: a comparison of library preparation methods for whole genome sequencing of Cryptococcus neoformans using Illumina HiSeq. *PLoS One,* 9**,** e113501.

RICHELDI, L. 2006. An update on the diagnosis of tuberculosis infection. *Am J Respir Crit Care Med,* 174**,** 736-42.

RUMORE, J., TSCHETTER, L., KEARNEY, A., KANDAR, R., MCCORMICK, R., WALKER, M., PETERSON, C.-L., REIMER, A. & NADON, C. 2018a. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157: H7 from the Canadian perspective. *BMC genomics,* 19**,** 1-13.

SABAT, A. J., VAN ZANTEN, E., AKKERBOOM, V., WISSELINK, G., VAN SLOCHTEREN, K., DE BOER, R. F., HENDRIX, R., FRIEDRICH, A. W., ROSSEN, J. W. A. & KOOISTRA-SMID, A. 2017. Targeted next-generation sequencing of the 16S-23S rRNA region for culture-independent bacterial identification - increased discrimination of closely related species. *Sci Rep,* 7**,** 3434.

SABER, M. M. & SHAPIRO, B. J. 2020. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom,* 6.

SAEB, A. T. M. 2018. Current Bioinformatics resources in combating infectious diseases. *Bioinformation,* 14**,** 31-35.

SALJE, H., LESSLER, J., MALJKOVIC BERRY, I., MELENDREZ, M. C., ENDY, T., KALAYANAROOJ, S., A-NUEGOONPIPAT, A., CHANAMA, S., SANGKIJPORN, S. & KLUNGTHONG, C. 2017. Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size. *Science,* 355**,** 1302-1306.

SAMPATH, R., HALL, T. A., MASSIRE, C., LI, F., BLYN, L. B., ESHOO, M. W., HOFSTADLER, S. A. & ECKER, D. J. 2007. Rapid identification of emerging infectious agents using PCR and electrospray ionization mass spectrometry. *Ann N Y Acad Sci,* 1102**,** 109-20.

SANDRIN, T. R. & DEMIREV, P. A. 2018. Characterization of microbial mixtures by mass spectrometry. *Mass Spectrometry Reviews,* 37**,** 321-349.

SAUER, S. & KLIEM, M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nat Rev Microbiol,* 8**,** 74-82.

SCHAUER, N., STEINHAUSER, D., STRELKOV, S., SCHOMBURG, D., ALLISON, G., MORITZ, T., LUNDGREN, K., ROESSNER-TUNALI, U., FORBES, M. G., WILLMITZER, L., FERNIE, A. R. & KOPKA, J. 2005. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett,* 579**,** 1332-7.

SCHNEIDER, M. V. & ORCHARD, S. 2011. Omics technologies, data and bioinformatics principles. *Bioinformatics for omics Data***,** 3-30.

SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-9.

SEGERMAN, B. 2020. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. *Front Cell Infect Microbiol,* 10**,** 527102.

SEIFERT, H., DOLZANI, L., BRESSAN, R., VAN DER REIJDEN, T., VAN STRIJEN, B., STEFANIK, D., HEERSMA, H. & DIJKSHOORN, L. 2005. Standardization and interlaboratory reproducibility assessment of pulsed-field gel electrophoresis-generated fingerprints of Acinetobacter baumannii. *Journal of Clinical Microbiology,* 43**,** 4328-4335.

SENES, C., SALDAN, N., COSTA, W., SVIDZINSKI, T. & OLIVEIRA, C. 2018. Identification of Fusarium oxysporum Fungus in Wheat Based on Chemical Markers and Qualitative GC-MS Test. *Journal of the Brazilian Chemical Society*.

SETH, B., DALAL, S. & KUMAR, R. 2019. Securing bioinformatics cloud for big data: Budding buzzword or a glance of the future. *Recent advances in computational intelligence.* Springer.

SHAMOUT, F., ZHU, T. & CLIFTON, D. A. 2021. Machine Learning for Clinical Outcome Prediction. *IEEE Rev Biomed Eng,* 14**,** 116-126.

SINGHAL, N., KUMAR, M., KANAUJIA, P. K. & VIRDI, J. S. 2015. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front Microbiol,* 6**,** 791.

SMITH, P. B., SNYDER, A. P. & HARDEN, C. S. 1995. Characterization of bacterial phospholipids by electrospray ionization tandem mass spectrometry. *Analytical chemistry,* 67**,** 1824-1830.

SOUVOROV, A., AGARWALA, R. & LIPMAN, D. J. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol,* 19**,** 153.

SPRATT, D. A. 2004. Significance of bacterial identification by molecular biology methods. *Endodontic topics,* 9**,** 5-14.

STEWART-IBARRA, A. M., RYAN, S. J., KENNESON, A., KING, C. A., ABBOTT, M., BARBACHANO-GUERRERO, A., BELTRAN-AYALA, E., BORBOR-CORDOVA, M. J., CARDENAS, W. B., CUEVA, C., FINKELSTEIN, J. L., LUPONE, C. D., JARMAN, R. G., MALJKOVIC BERRY, I., MEHTA, S., POLHEMUS, M., SILVA, M. & ENDY, T. P. 2018. The Burden of Dengue Fever and Chikungunya in Southern Coastal Ecuador: Epidemiology, Clinical Presentation, and Phylogenetics from the First Two Years of a Prospective Study. *Am J Trop Med Hyg,* 98**,** 1444-1459.

SUTTON, S. 2007. How do you decide which microbial identification system is best. *microbiology,* 585**,** 210-8336.

TAYLOR, T. A. & UNAKAL, C. G. 2021. Staphylococcus aureus. *StatPearls [Internet]*.

TENG, J. L. L., YEUNG, M. L., CHAN, E., JIA, L., LIN, C. H., HUANG, Y., TSE, H., WONG, S. S. Y., SHAM, P. C., LAU, S. K. P. & WOO, P. C. Y. 2017. PacBio But Not Illumina Technology Can Achieve Fast, Accurate and Complete Closure of the High GC, Complex Burkholderia pseudomallei Two-Chromosome Genome. *Front Microbiol,* 8**,** 1448.

TSHIBANGU-KABAMBA, E., NGOMA-KISOKO, P. D. J., TUAN, V. P., MATSUMOTO, T., AKADA, J., KIDO, Y., TSHIMPI-WOLA, A., TSHIAMALA-KASHALA, P., AHUKA-MUNDEKE, S. & MUMBA NGOY, D. 2020. Next-generation sequencing of the whole bacterial genome for tracking molecular insight into the broad-spectrum antimicrobial resistance of Helicobacter pylori clinical isolates from the Democratic Republic of Congo. *Microorganisms,* 8**,** 887.

VAN BELKUM, A., VAN LEEUWEN, W., KAUFMANN, M. E., COOKSON, B., FOREY, F., ETIENNE, J., GOERING, R., TENOVER, F., STEWARD, C. & O'BRIEN, F.

1998. Assessment of resolution and intercenter reproducibility of results of genotyping Staphylococcus aureus by pulsed-field gel electrophoresis of Sma I macrorestriction fragments: a multicenter study. *Journal of Clinical Microbiology,* 36**,** 1653-1659.

VAN EMBDEN, J., CAVE, M. D., CRAWFORD, J. T., DALE, J., EISENACH, K., GICQUEL, B., HERMANS, P., MARTIN, C., MCADAM, R. & SHINNICK, T. 1993. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *Journal of clinical microbiology,* 31**,** 406-409.

VAN GOETHEM, N., DESCAMPS, T., DEVLEESSCHAUWER, B., ROOSENS, N. H., BOON, N. A., VAN OYEN, H. & ROBERT, A. 2019. Status and potential of bacterial genomics for public health practice: a scoping review. *Implementation Science,* 14**,** 1-16.

VÁRADI, L., LUO, J. L., HIBBS, D. E., PERRY, J. D., ANDERSON, R. J., ORENGA, S. & GROUNDWATER, P. W. 2017. Methods for the detection and identification of pathogenic bacteria: past, present, and future. *Chemical Society Reviews,* 46**,** 4818-4832.

VARLAMOV, D. A., BLAGODATSKIKH, K. A., SMIRNOVA, E. V., KRAMAROV, V. M. & IGNATOV, K. B. 2020. Combinations of PCR and Isothermal Amplification Techniques Are Suitable for Fast and Sensitive Detection of SARS-CoV-2 Viral RNA. *Front Bioeng Biotechnol,* 8**,** 604793.

WANG, P., PANG, S., ZHANG, H., FAN, M. & HE, L. 2016. Characterization of Lactococcus lactis response to ampicillin and ciprofloxacin using surface-enhanced Raman spectroscopy. *Analytical and bioanalytical chemistry,* 408**,** 933-941.

WICK, R. R., JUDD, L. M., GORRIE, C. L. & HOLT, K. E. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol,* 13**,** e1005595.

WOOLEY, J. C., GODZIK, A. & FRIEDBERG, I. 2010. A primer on metagenomics. *PLoS Comput Biol,* 6**,** e1000667.

WORDSWORTH, S., DOBLE, B., PAYNE, K., BUCHANAN, J., MARSHALL, D. A., MCCABE, C. & REGIER, D. A. 2018. Using "Big Data" in the Cost-Effectiveness Analysis of Next-Generation Sequencing Technologies: Challenges and Potential Solutions. *Value Health,* 21**,** 1048-1053.

YANG, S. & ROTHMAN, R. E. 2004. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *The Lancet infectious diseases,* 4**,** 337-348.

YEUNG, P., SANDERS, M., KITTS, C. L., CANO, R. & TONG, P. S. 2002. Species-specific identification of commercial probiotic strains. *Journal of Dairy Science,* 85**,** 1039-1051.

ZHANG, Y., ODIWUOR, N., XIONG, J., SUN, L., NYARUABA, R. O., WEI, H., ET AL. 2020. Rapid molecular detection of SARS-CoV-2 (COVID-19) Virus RNA using colorimetric LAMP. *MedRxiv (Preprint).*

ZHAO, F. & BAJIC, V. B. 2015. The Value and Significance of Metagenomics of Marine Environments. Preface. *Genomics Proteomics Bioinformatics,* 13**,** 271-4.

ZHAO, X., LIN, C.-W., WANG, J. & OH, D. H. 2014. Advances in rapid detection methods for foodborne pathogens. *Journal of microbiology and biotechnology,* 24**,** 297-312.

ZOU, J., HUSS, M., ABID, A., MOHAMMADI, P., TORKAMANI, A. & TELENTI, A. 2019. A primer on deep learning in genomics. *Nat Genet,* 51**,** 12-18.

# Chapter 2

## Genomic Evolution of *Staphylococcus aureus* During Artificial and Natural Colonization of the Human Nose

Manisha Goyal[1], Fabien Javerliat[2], Mattia Palmieri[1], Caroline Mirande[2], Willem van Wamel[3], Mehri Tavakol[3], Nelianne J. Verkaik[3] and Alex van Belkum[1]*

[1]Data Analytics Unit, bioMérieux, La Balme-les-Grottes, France

[2]Microbiology R&D, bioMérieux, La Balme-les-Grottes, France

[3]Department of Medical Microbiology and Infectious Diseases, Erasmus University Medical Center, Rotterdam, Netherlands

## Abstract

*Staphylococcus aureus* can colonize the human vestibulum nasi for many years. It is unknown whether and, how *S. aureus* adapts to this ecological niche during colonization. We determined the short (1 and 3 months) and mid-term (36 months) genomic evolution of *S. aureus* in natural carriers and artificially colonized volunteers. Eighty-five *S. aureus* strains were collected from 6 natural carriers during 3 years and 6 artificially colonized volunteers during 1 month. Multi-locus sequence typing (MLST) and single nucleotide polymorphism (SNP) analysis based on whole-genome sequencing (WGS) were carried out. Mutation frequencies within resident bacterial populations over time were quantified using core genome SNP counts (comparing groups of genomes) and pairwise SNP divergence assessment (comparing two genomes from strains originating from one host and sharing identical MLST). SNP counts (within 1–3 months) in all naturally colonizing strains varied from 0 to 757 (median 4). These strains showed random and independent patterns of pairwise SNP divergence (0 to 44 SNPs, median 7). When the different core genome SNP counts over a period of 3 years were considered, the median SNP count was 4 (range 0–26). Host-specific pairwise SNP divergence for the same period ranged from 9 to 57 SNPs (median 20). During short term artificial colonization the mutation frequency was even lower (0–7 SNPs, median 2) and the pairwise SNP distances were 0 to 5 SNPs (median 2). Quantifying mutation frequencies is important for the longitudinal follow-up of epidemics of infections and outbreak management. Random pattern of pairwise SNP divergence between the strains isolated from single carriers suggested that the WGS of multiple colonies is necessary in this context. Over periods up to 3 years, maximum median core genome SNP counts and SNP divergence for the strains studied were 4 and 20 SNPs or lower. During artificial colonization, where median core genome SNP and pairwise SNP distance scores were 2, there is no early stage selection of different genotypes. Therefore, we suggest an epidemiological cut off value of 20 SNPs as a marker of *S. aureus* strain identity during studies on nasal colonization and also outbreaks of infection.

## Introduction

Extensive use of antibiotics in the environment and the clinical domain contributes toward the emergence of (multi-)drug resistant bacterial pathogens. This has become a global threat (Roca et al., 2015)). *Staphylococcus aureus* (*S. aureus*) is among the bacterial species associated with increasing drug resistance, morbidity, invasive disease, and mortality in humans as well as animals (Chambers and Deleo, 2009, Li and Webster, 2018, Schmidt et al., 2015). *S. aureus* is a common opportunistic human pathogen identified most often on the nasal epithelium, About 30–50% of healthy individuals are persistently colonized (Wertheim et al., 2005). *S. aureus* causes a large variety of community as well as hospital-acquired infections. These include deep abscesses, endocarditis, osteomyelitis, pneumonia, and bloodstream infections (Foster and Höök, 1998, Rasigade and Vandenesch, 2014, Taylor and Unakal, 2018). *S. aureus* nasal carriage is a risk factor for the development of staphylococcal infections. Adherence to the human nasal epithelial cells is a prerequisite for *S. aureus* colonization and initiation of infection (Roche et al., 2003). The prevalence of non-symptomatic colonization with methicillin resistant *S. aureus* strains in the open United States population escalated from 0.8 to 1.5% over recent years (Gorwitz et al., 2008).

Colonization begins with the interaction between nasal epithelial ligands and bacterial receptors often cataloged as microbial surface components recognizing adhesive matrix molecules (MSCRAMMS) (Foster and Höök, 1998, Ghasemian et al., 2015). During colonization *S. aureus* expresses adherence genes (clf B, isdA, fnbA, eap, sceD, oatA, and atlA) and several immune-modulating genes (e.g., sak, chp, spa, and scn) (Baur et al., 2014, Burian et al., 2010). Host factors and local microbiota can affect the adhesion and colonization properties of *S. aureus* as well (Emonts et al., 2007, Frank et al., 2010, Ruimy et al., 2010).

During colonization, *S. aureus* secretes a number of immune-modulating proteins. Staphylococcal complement inhibitor (SCIN), encoded by the scn gene, can efficiently protect *S. aureus* by inhibiting the innate immune response mediated by human neutrophils. SCIN and other immune modulating proteins are encoded on the immune evasion cluster (IEC) (Goerke et al., 2006). The scn gene was identified as a conserved one being present in all IEC (van Wamel et al., 2006). To test the role and stability of IEC human artificial inoculation was performed using isolates with and without IEC. It was concluded that IEC may not play a significant role in adherence but it did display an essential role in propagation and long term survival (Verkaik et al., 2011).

We have here used whole genome sequencing (WGS) to quantify the mutational changes occurring in *S. aureus* strains during natural and artificial nasal colonization during periods ranging between 1 and 36 months. The numbers of human volunteers and hence the overall number of *S. aureus* nasal isolates are limited due to the technical and logistic complexity of the studies involved (Verkaik et al., 2011). In addition, studies involving colonization of human volunteers have to follow extensive ethical procedures and protocols. We applied bio-informatics approaches to assign MLST types and to detect genetic variation at the single nucleotide polymorphism (SNP) level. Moreover, we analyzed selective presence of virulence factors for all strains.

## Materials and Methods

### Description of the Strain Collection

*Staphylococcus aureus* strain collection was carried out as described earlier (Verkaik et al., 2011) at Erasmus Medical Center (Rotterdam, Netherlands). Naturally colonizing strains were isolated from nasal swab cultures from healthy persistent carriers who were positive for *S. aureus* at five culture moments over a time interval of 3 months in both 2007 and 2010. Artificially colonizing strains were collected from the human volunteers inoculated with *S. aureus* strain NCTC 8325-4 with or without IEC and follow-up cultures were performed in 2008 (days 1, 2, 3, 4, 7, 14, 21, and 28 after inoculation). The latter strains were susceptible to all common antibiotics and were free from staphylococcal toxin genes (Wertheim et al., 2008, Williams et al., 1997). A review of all strains sequenced is provided in Supplementary Table 1.

### *S. aureus* Genome Sequences

Isolates were sequenced by WGS (Illumina HiSeq 2000 platform). Raw reads were assembled using the A5 MiSeq-20140604 assembler. Datasets for strains cultured in 2007 and 2010 contained 35 and 22 isolates, respectively, involving natural nasal colonization in 6 persistent carriers. The dataset from 2008 (28 isolates) was collected for strains from 6 different volunteers artificially colonized with *S. aureus* strain NCTC 8325-4. DNA isolation was performed for up to three colonies from each culture to define their genotypic stability at different point of times during short term as well longer term nasal carriage (see

Supplementary Table 1. The sequences obtained from the 1 to 3 independent colonies taken for some of the individual strains were analyzed independently by the bioinformatics tools applied. Using bioinformatics tools as BioNumerics (Applied Maths, bioMérieux, Belgium), kSNP3 (Computations/Global Security, Lawrence Livermore National Laboratory, Livermore, CA, United States and Bellingham Research Institute, Bellingham, WA, United States), and Abricate (Torsten Seemann, University of Melbourne, Australia), all genomes were analyzed extensively.

**MLST Typing and SNP Detection**

To understand the genetic diversity of all the isolates multi-locus sequence typing (MLST) was performed using BioNumerics1. The MLST method is known to have a higher discriminatory power for *S. aureus* strains than PFGE (Peacock et al., 2002). For classical MLST typing seven housekeeping genes and their various alleles were used to define strain relatedness (Jolley et al., 2018). A phylogenetic tree was constructed by executing the Linux based stand-alone source code of kSNP3 (Gardner et al., 2015), which identified core genome SNP counts and provided a consensus parsimony phylogenetic tree. The kmer size was set to 19, the optimum size estimated by the kSNP3 utility program Kchooser (Gardner et al., 2015). Pairwise SNP distances between later stage isolates as compared to early stage isolates from each individual were calculated to define mutation over time. The python script kSNPdist was used to calculate the pairwise SNP divergence between all *S. aureus* isolates. kSNP3 and kSNPdist executables for OS X and Linux are freely available at https://sourceforge.net/projects/ksnp/.

**Resistance and Virulence Gene Identification**

All the genomes were screened for the presence of 40 known and putative virulence genes (Shukla et al., 2010) (enterotoxin genes, exotoxin genes, leucocidin genes, hemolysin genes, surface protein genes, and putative virulence genes) and the *S. aureus* antibiotic resistance genes available in the ResFinder database4. Those 40 genes are grouped as classical staphylococcal. The Linux-based command line tool known as Abricate was downloaded5 to perform additional mass screening for antimicrobial resistance or virulence genes. All the identified resistance and virulence genes in the dataset were summarized in

Supplementary Table 3. Additionally, in silico-based mapping of the scn gene using BioNumerics was carried out to determine the presence of IEC (van Wamel et al., 2006).

## Results

### Quality Testing of Genome Datasets

Genome sizes varied from 2,647 to 2,827 Kilo base (kb). The average number of contigs generated per genome was 64 contigs (ranging from 40 to 315 contigs). The average N50 contig length was 171778 bp (Supplementary Table 2). Isolates (and hence their genomes) from a single individual are expected to be part of a single clade as predicted by the MLST data and phylogenetic clustering (Figures 2- 1, 2).

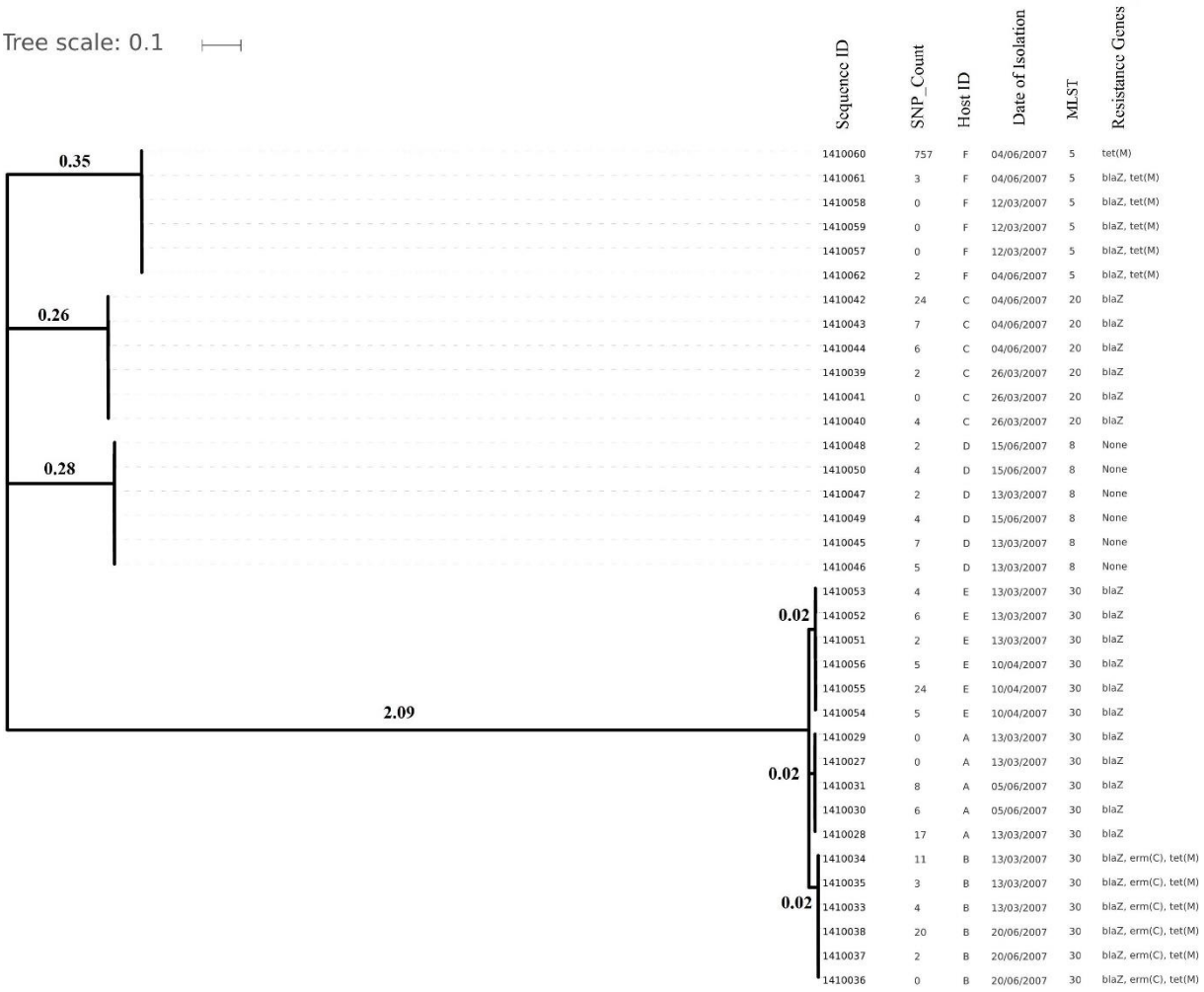



***Figure 2-1:*** *Phylogenetic tree depicting clustering on the basis of core SNP count ranges from 0 to 757 SNPs (median 4 SNPs) in all the Staphylococcus aureus strains colonized during 3 months (2007 subgroup) of follow up along with their date of isolation, persistent carriers from which they have isolated after maximum three cultural moments, their sequence*

*type and resistance genes. Note that all isolates are clustered together on the basis of the original individual they were cultured from.*

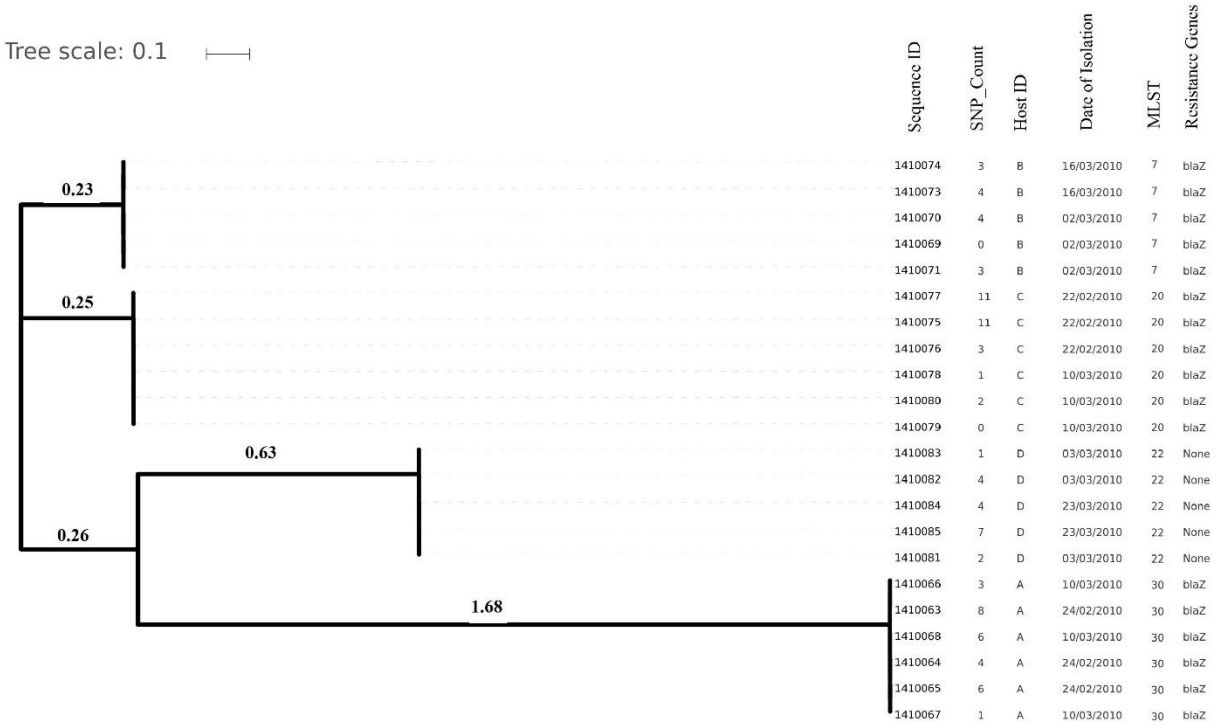



***Figure 2-2:*** *Evolutionary relationship on the basis of core genome SNP counts detected (range 0 to 11 SNPs) in S. aureus strains colonized and isolated during 1 month (2010 subgroup) along with the date of their isolation, the host from which they have isolated, MLST and resistance genotype. Isolates from the same host are clustered together showing their higher strain relatedness.*

**Short Term Evolution (1–3 Months) in Naturally Colonizing *S. aureus* Strains**

*Staphylococcus aureus* isolates from 2007 and 2010 (35 and 22 genomes, respectively) from carriers A to F were analyzed for short term genomic changes, over 3 months (in 2007), and 1 month (in 2010). ST30 (2007: 47%; 2010: 30%) and ST20 (2007: 17.60%; 2010: 30%) were found to be the dominant MLST types followed by ST8 and ST5 (18% each) in 2007 and ST22 and ST7 (20% each) in the 2010 subgroups (Table 2- 1). Over the period of 3 years some of the strains were replaced by different sequence type strains within a same carrier. For instance isolates from carrier B and D in 2007 were ST30 and ST8 but in 2010, isolates from the same carries were ST7 and ST22, respectively. These strains were not included for longer term pairwise SNP divergence analysis (Table 2- 1).

***Table 2-1:*** *Pairwise SNP distances identified between all the early and later stages isolates (according to their isolation date) among the S. aureus strains of subgroup 2007 and 2010 independently from each persistent nasal carrier.*

| | 2007 | | | | 2010 | | | |
|---|---|---|---|---|---|---|---|---|
| **Persistent carrier ID** | **Begin (Seq. ID)** | **Isolate after 3 months (Seq. ID)** | **No. of SNP differences** | **MLST** | **Begin (Seq. ID)** | **Isolate after 1 month (Seq. ID)** | **SNP differences** | **MLST** |
| A | 1410027 | 1410030 | 10 | 30 | 1410063 | 1410066 | 4 | 30 |
| | | 1410031 | 11 | | | 1410067 | 5 | |
| | | – | – | | | 1410068 | 7 | |
| | 1410028 | 1410030 | 11 | | 1410064 | 1410066 | 4 | |
| | | 1410031 | 10 | | | 1410067 | 3 | |
| | | – | – | | | 1410068 | 3 | |
| | 1410029 | 1410030 | 11 | | 1410065 | 1410066 | 3 | |
| | | 1410031 | 13 | | | 1410067 | 3 | |
| | | – | – | | | 1410068 | 3 | |
| B | 1410033 | 1410036 | 7 | 30 | 1410069 | 1410073 | 11 | 7 |
| | | 1410037 | 7 | | | 1410074 | 12 | |
| | | 1410038 | 10 | | | – | – | |
| | 1410034 | 1410036 | 17 | | 1410070 | 1410073 | 10 | |
| | | 1410037 | 17 | | | 1410074 | 11 | |
| | | 1410038 | 20 | | | – | – | |
| | 1410035 | 1410036 | 7 | | 1410071 | 1410073 | 13 | |
| | | 1410037 | 7 | | | 1410074 | 14 | |
| | | 1410038 | 10 | | | – | – | |
| C | 1410039 | 1410042 | 43 | 20 | 1410075 | 1410078 | 7 | 20 |
| | | 1410043 | 16 | | | 1410079 | 7 | |
| | | 1410044 | 16 | | | 1410080 | 9 | |
| | 1410040 | 1410042 | 44 | | 1410076 | 1410078 | 3 | |
| | | 1410043 | 18 | | | 1410079 | 2 | |
| | | 1410044 | 18 | | | 1410080 | 2 | |
| | 1410041 | 1410042 | 43 | | 1410077 | 1410078 | 8 | |
| | | 1410043 | 15 | | | 1410079 | 7 | |
| | | 1410044 | 15 | | | 1410080 | 8 | |
| D | 1410045 | 1410048 | 25 | 8 | 1410081 | 1410084 | 2 | 22 |
| | | 1410049 | 3 | | | 1410085 | 3 | |
| | | 1410050 | 26 | | 1410082 | 1410084 | 1 | |
| | 1410046 | 1410048 | 27 | | | 1410085 | 4 | |
| | | 1410049 | 3 | | 1410083 | 1410084 | 2 | |
| | | 1410050 | 27 | | | 1410085 | 4 | |
| | 1410047 | 1410048 | 24 | | | | | |
| | | 1410049 | 3 | | | | | |
| | | 1410050 | 26 | | | | | |
| E | 1410051 | 1410054 | 2 | 30 | | | | |
| | | 1410055 | 1 | | | | | |
| | | 1410056 | 0 | | | | | |
| | 1410052 | 1410054 | 6 | | | | | |
| | | 1410055 | 22 | | | | | |
| | | 1410056 | 5 | | | | | |
| | 1410053 | 1410054 | 5 | | | | | |
| | | 1410055 | 22 | | | | | |
| | | 1410056 | 3 | | | | | |
| F | 1410057 | 1410060 | 9 | 5 | | | | |
| | | 1410061 | 2 | | | | | |
| | | 1410062 | 0 | | | | | |
| | 1410058 | 1410060 | 9 | | | | | |
| | | 1410061 | 2 | | | | | |
| | | 1410062 | 0 | | | | | |
| | 1410059 | 1410060 | 8 | | | | | |
| | | 1410061 | 2 | | | | | |
| | | 1410062 | 0 | | | | | |

*Multilocus sequence typing was assigned to all the isolates of each host. From one carrier, multiple colonies were selected and sequenced to count the number of SNP differences between different colonies. Exceptionally from carrier A strain id 1410032 was shifted to ST22 and isolate 1410072 was contaminated thus not included in SNP analyses.*

Core genome SNP counts for the genomes of all the strains collected in 2007 and 2010 ranged from 0 to 757 SNPs (median 4 SNPs) and 0 to 11 SNPs (median 3.5 SNPs), respectively (Figures 2- 1, 2). We observed a small pairwise SNP distance between all the early and the later stage isolates within a carrier (all carriers pooled, 2007 median number of SNP divergence was 10 and in 2010 median SNP distance was 4 (Table 2- 1). The maximum number of pairwise SNP differences calculated for the genomes of the isolates of carrier C ranged from 15 to 44 SNPs followed by 3 to 27 SNPs in strains from carrier D, 0 to 22 in strains from E, 7 to 20 in strains from B, 10 to 13 in strains from A and 0 to 9 SNPs in strains from F (Table 2- 1). Paired SNP differences were also calculated for strains from subgroup 2010 illustrating the highest ranges (10–14) among strains from host B followed by 2 to 9 SNPs in strains from C, 3 to 7 in strains from A, and 1 to 4 in strains from individual D (Table 2- 1). On an individual basis, the pattern of pairwise SNP differences is relatively random between the isolates from early and later stages of colonization.

**Longer Term Evolution (3 Years) in Naturally Colonizing *S. aureus* Strains**

Evolutionary analysis over a period of 3 years (2007–2010) could only be done for the isolates from two persistent carriers, A, and C. In these carriers the MLST type remained unchanged over time, suggesting persistent colonization with the same strains (Table 2- 1). All isolates of carrier A and C from both 2007 and 2010 were analyzed for the presence of core genome SNPs which ranged from 0 to 26 SNPs (median 4 SNPs) (Figure 2- 3). Host specific pairwise SNP differences between the isolates dating 2007 and 2010 for both carriers A and C individually were 9–33 SNPs (median 19) and 15-57 SNPs (median 24), respectively (Figure 2- 4). All strains from one carrier showed random distribution of SNPs; e.g., SNP distances between strains 1410027 and 1410029 versus the later stage strain 1410066 were 9 and 11 SNPs (Figure 2- 4). This demonstrated that genomic evolution was random and none of the SNPs were fixed genetically over time.

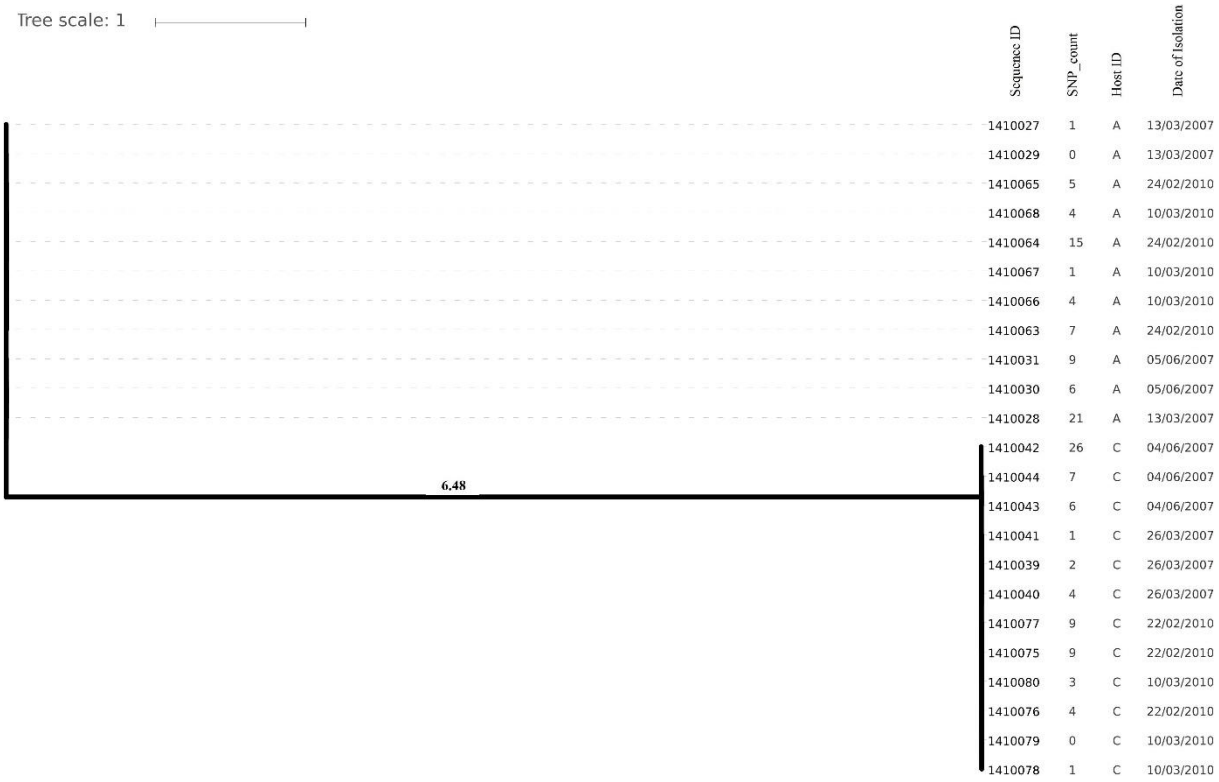



***Figure 2-3:*** *Phylogenetic Tree showing longer term (3 yeras) diversity and relatedness of S. aureus strains on the bases of core genome SNP counts ranged from 0 to 26 SNPs in all the isolates from two nasal carriage individuals (A and C) for both the years 2007 and 2010.*

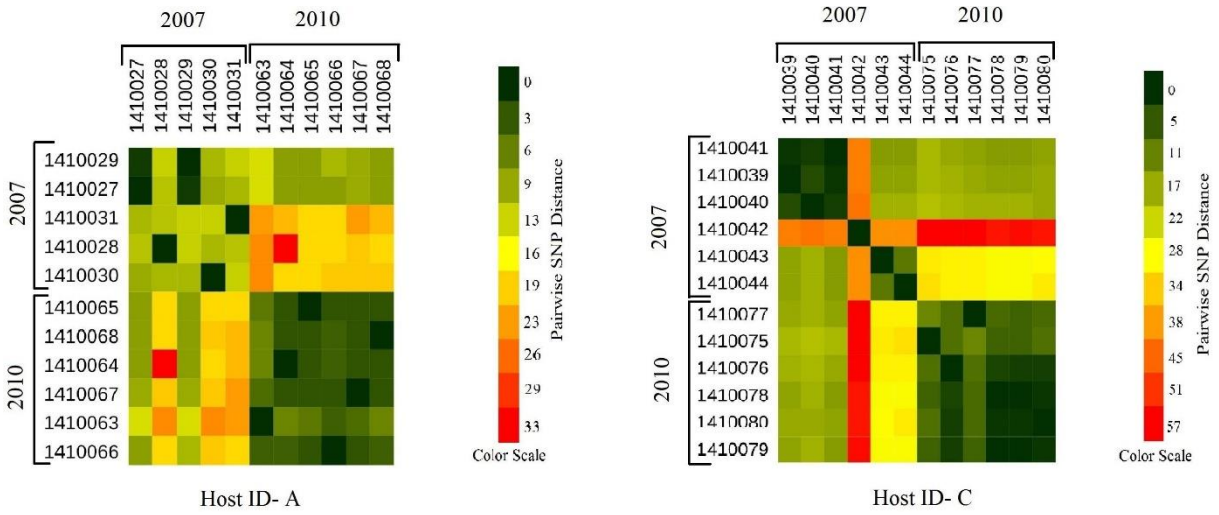



***Figure 2-4:*** *Heat maps showing the host specific pairwise SNP (longer term) among all the early (2007) and later stage (2010) isolates of carrier A and carrier C individually with the color range of dark green (least SNP divergence) to red (higher SNP divergence). In both the hosts A and C pairwise SNP distances between the isolates of 2007 and 2010 datasets are visibly higher (from yellow to red boxes) than that of within the dataset itself (from dark green to light green boxes) with one exceptional isolate 1410042 in carrier C which showed higher pairwise SNP divergence within its dataset (orange boxes) as well as with the isolates of 2010 dataset (red boxes).*

**Mutational Analyses of Strains From Artificially Colonized Humans**

All strains were of ST8. No considerable identity was observed with resistance genes (Supplementary Table 3) from the database which was in agreement with the pan-susceptibility of the isolates. The overall core genome SNP counts among the isolates ranged from 0 to 7 SNPs (average 2) (Figure 2- 5). The maximum range of pairwise SNP distances between the isolates within a host was 0 to 5 SNPs (median SNP distance 2) after 28 days of colonization in *S. aureus* nasal carriers (Table 2- 2).

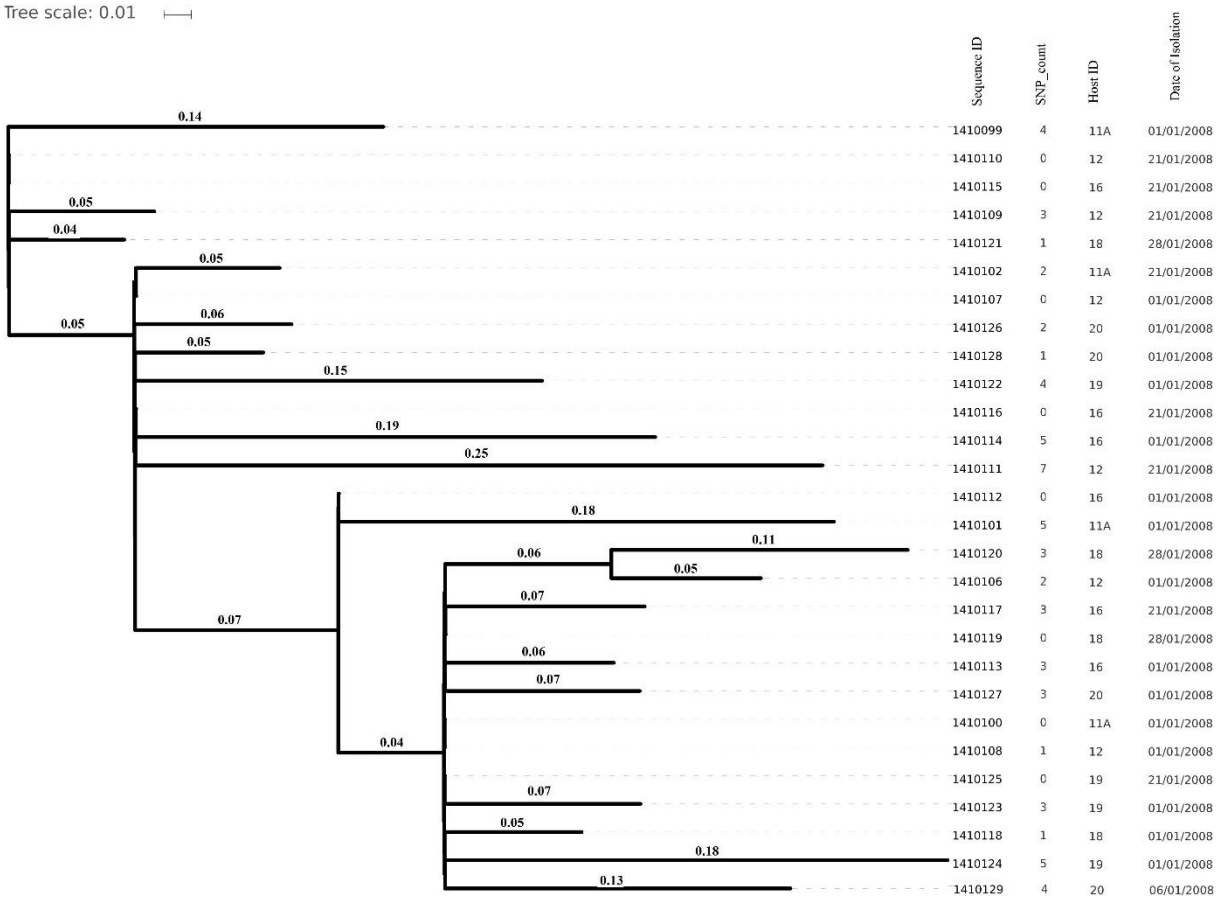



***Figure 2-5:*** *Core genome SNP counts based phylogenetic tree illustrating the close resemblance among the genomes isolated from artificially inoculated S. aureus nasal carriers in 2008. Core genome SNP counts here ranged from 0 to 7 core SNPs and each cluster is showing random collection of the strains irrespective of their specific host depicted very less genomic evolution (in 1 month) in artificially colonizing strains.*

***Table 2-2:*** *Pairwise SNP distances found in artificially colonizing strains isolated during short term colonization in different individuals.*

| Host ID | 2008 (28 days) | | SNP differences |
|---|---|---|---|
| | Begin | End | |
| 11A | 1410099 | 1410102 | 4 |
| | 1410100 | | 1 |
| | 1410101 | | 4 |
| 12 | 1410106 | 1410109 | 3 |
| | | 1410110 | 2 |
| | | 1410111 | 5 |
| | 1410107 | 1410109 | 2 |
| | | 1410110 | 1 |
| | | 1410111 | 4 |
| | 1410108 | 1410109 | 2 |
| | | 1410110 | 0 |
| | | 1410111 | 3 |
| 16 | 1410112 | 1410115 | 0 |
| | | 1410116 | 0 |
| | | 1410117 | 0 |
| | 1410113 | 1410115 | 1 |
| | | 1410116 | 1 |
| | | 1410117 | 1 |
| | 1410114 | 1410115 | 3 |
| | | 1410116 | 2 |
| | | 1410117 | 3 |
| 18 | 1410118 | 1410119 | 1 |
| | | 1410120 | 4 |
| | | 1410121 | 2 |
| 19 | 1410122 | 1410125 | 2 |
| | 1410123 | | 1 |
| | 1410124 | | 2 |
| 20 | 1410126 | 1410129 | 3 |
| | 1410127 | | 3 |
| | 1410128 | | 3 |

Fourteen virulence genes (*sea, hla, hlb, hld, hlgB, clfA, clfB, fnbA, fnbB, icaA, sdrC, sdrD, sdrE, and tsst-1*) were identified in the current sequence dataset (Supplementary Table 3). The virulence factor *fnbA* was not found in isolates from host B and was also missing in one of the isolates from carrier F (1410060). Two strains (1410054 and 1410055) were shown to have acquired the *cna* gene during colonization of host E (Supplementary Table 3). Absence of the *scn* gene corroborating the complete lack of IEC in artificially colonized strains (Supplementary Table 3).

## Discussion

In the present work, we have studied the evolutionary patterns in nasal *S. aureus* strains to better understand their local adaptive behavior and mutational frequency. Low core genomes SNP values among all the genomes defines the significant strain relatedness witnessed in this study. This is experimentally supported by the outcomes of previous research (Ankrum and Hall, 2017) where *S. aureus* strains with <71 SNP differences were considered as non-discriminate. Similar findings by (Golubchik et al., 2013) suggested that SNP divergence with in a host varied from ~ 0 to 27 SNPs among host specific isolates. In our study, one isolate from host C (1410042) was showing an exceptionally high SNP divergence value for which we have no clear explanation (Figure 2- 5 and Table 2- 1). Phylogenetic trees (Figures 2- 1, 2, 3, 5) were constructed on the basis of core genome SNPs identified within strains from all individual hosts showing different numbers of mutations as compared to their pairwise frequency of SNP divergence. The level of diversity (SNP divergence) within the hosts was consistently lower than that detected between different hosts and of same MLST type (Golubchik et al., 2013).

Prior studies tried to assess the number of SNPs accumulating over time, but mostly under selective conditions. (Rouard et al., 2018) calculated that during selection for linezolid resistance an expected 17–93 mutations should accumulate per genome per year. A more global calculation using a significantly larger strain collection resulted in average number of less than 10 SNPs per genome per year (Harris et al., 2010). (Ankrum and Hall, 2017) came up with figures around 70 SNPs per year. Obviously, the discussion on epidemiological SNP cut off values defining identity (or not) or close relatedness between clinical isolates have not been finalized yet. On the basis of this study, we suggest a median SNP cut off 20 SNPs. Although in our study limited numbers of individuals are included, a high number of strains per individual were included to thoroughly study mutation frequency over time. Our suggested cut off can be used to identify *S. aureus* strains as identical or not in outbreak management.

Nasal colonization with strains carrying virulence determinants such as *fibronectin* (*fnb*) and *collagen adhesions* (*cna*) may represent risk for subsequent invasive infections in carriers (Nashev et al., 2004, Peacock et al., 2002).

We acknowledge that our study is likely to be underpowered: limited numbers of individuals were able to take part in these long-lasting and logistically complicated studies.

Still, the artificial inoculation model is a unique feature of this study. So far and except for our own work, very few studies have been done using artificial inoculation in humans (Cole et al., 2018). On the other hand, epidemiological studies usually take place in similar time frames as used here. The mutation frequency we observe here during weeks and months will be well aligned with those occurring during active outbreaks since these mostly also span weeks rather than months.

## Conclusion

Median core genome SNP counts and pairwise SNP divergence for all the strains studied here were always lower than 20 over periods up to 3 years of evolution in individual carriers. During artificial colonization, where median core genome SNP, and pairwise SNP distance scores were 2, there is no early stage selection of different genotypes. In addition, during stable long(er) term colonization (up to 3 years) the number of accumulating SNPs was low as well. We here suggest an epidemiological median cut off value of 20 SNPs as a marker of *S. aureus* strain identity during outbreaks of infection. Random pattern of pairwise SNP divergence between the strains isolated from single carrier suggested that the WGS of multiple colonies is necessary for outbreak infection analysis.

## Data Availability

The datasets for this manuscript are not publicly available because we are still in the process of submitting data on NCBI. Requests to access the datasets should be directed to manishagoyal.rama@gmail.com.

## Author Contributions

WvW, NV, and AvB conceived the study. MT conducted the microbiological experimentation for *S. aureus* strains. MG, FJ, MP, and CM carried out the whole genome sequencing studies. MG interpreted the sequence data and wrote the first version of the manuscript. All authors discussed the results and edited the manuscript.

## Funding

This work was funded by bioMérieux, France, the Erasmus University Medical Center, Netherlands, and the ViBrANT (ITN project Marie Skłodowska-Curie Grant Agreement No. 765042 funded by the European Union).

## Conflict of Interest Statement

AvB, MP, CM, FJ, and MG were employees of bioMérieux, a company designing, developing, and marketing tests in the domain of infectious diseases. The company was not involved in the design of the current review and the opinions expressed are those of the authors and may be different from formal company opinions and policies. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Acknowledgments

We acknowledge the volunteers who were inoculated with *S. aureus* strains to carry out the study.

## Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01525/full#supplementary-material

## References

ANKRUM, A. & HALL, B. G. 2017. Population Dynamics of Staphylococcus aureus in Cystic Fibrosis Patients To Determine Transmission Events by Use of Whole-Genome Sequencing. *J Clin Microbiol,* 55**,** 2143–2152.

BAUR, S., RAUTENBERG, M., FAULSTICH, M., GRAU, T., SEVERIN, Y., UNGER, C., HOFFMANN, W. H., RUDEL, T., AUTENRIETH, I. B. & WEIDENMAIER, C. 2014. A nasal epithelial receptor for Staphylococcus aureus WTA governs adhesion to epithelial cells and modulates nasal colonization. *PLoS Pathog,* 10**,** e1004089.

BURIAN, M., WOLZ, C. & GOERKE, C. 2010. Regulatory adaptation of Staphylococcus aureus during nasal colonization of humans. *PLoS One,* 5**,** e10040.

CHAMBERS, H. F. & DELEO, F. R. 2009. Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nat Rev Microbiol,* 7**,** 629-41.

COLE, A. L., SCHMIDT-OWENS, M., BEAVIS, A. C., CHONG, C. F., TARWATER, P. M., SCHAUS, J., DEICHEN, M. G. & COLE, A. M. 2018. Cessation from Smoking Improves Innate Host Defense and Clearance of Experimentally Inoculated Nasal Staphylococcus aureus. *Infect Immun,* 86.

EMONTS, M., UITTERLINDEN, A. G., NOUWEN, J. L., KARDYS, I., DE MAAT, M. P. M., MELLES, D. C., WITTEMAN, J., DE JONG, P. T. V. M., VERBRUGH, H. A., HOFMAN, A., HERMANS, P. W. M. & A., V. B. 2007. Host Polymorphisms in Interleukin 4, Complement Factor H, and C-Reactive Protein Associated with Nasal Carriage of Staphylococcus aureus and Occurrence of Boils. *The Journal of Infectious Diseases,* 197**,** 1244–1253.

FOSTER, T. M. & HÖÖK, M. 1998. Surface protein adhesins of Staphylococcus aureus. *TRENDS IN MICROBIOLOGY,* 6**,** 484- 488.

FRANK, D. N., FEAZEL, L. M., BESSESEN, M. T., PRICE, C. S., JANOFF, E. N. & PACE, N. R. 2010. The human nasal microbiota and Staphylococcus aureus carriage. *PLoS One,* 5**,** e10598.

GARDNER, S. N., SLEZAK, T. & HALL, B. G. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics,* 31**,** 2877-8.

GHASEMIAN, A., PEERAYEH, S. N., BAKHSHI, B. & MIRZAEE, M. 2015. The Microbial Surface Components Recognizing Adhesive Matrix Molecules (MSCRAMMs) Genes among Clinical Isolates of Staphylococcus aureus from Hospitalized Children. *Iran J Pathol.* , 10**,** 258 - 264.

GOERKE, C., WIRTZ, C., FLUCKIGER, U. & WOLZ, C. 2006. Extensive phage dynamics in Staphylococcus aureus contributes to adaptation to the human host during infection. *Mol Microbiol,* 61**,** 1673-85.

GOLUBCHIK, T., BATTY, E. M., MILLER, R. R., FARR, H., YOUNG, B. C., LARNER-SVENSSON, H., FUNG, R., GODWIN, H., KNOX, K., VOTINTSEVA, A., EVERITT, R. G., STREET, T., CULE, M., IP, C. L., DIDELOT, X., PETO, T. E., HARDING, R. M., WILSON, D. J., CROOK, D. W. & BOWDEN, R. 2013. Within-host evolution of Staphylococcus aureus during asymptomatic carriage. *PLoS One,* 8**,** e61319.

GORWITZ, R. J., KRUSZON-MORAN, D., MCALLISTER, S. K., MCQUILLAN, G., MCDOUGAL, L. K., FOSHEIM, G. E., JENSEN, B. J., KILLGORE, G., TENOVER, F. C. & KUEHNERT, M. J. 2008. Changes in the prevalence of nasal colonization with Staphylococcus aureus in the United States, 2001-2004. *J Infect Dis,* 197**,** 1226-34.

HARRIS, S. R., FEIL, E. J., HOLDEN, M. T., QUAIL, M. A., NICKERSON, E. K., CHANTRATITA, N., GARDETE, S., TAVARES, A., DAY, N., LINDSAY, J. A., EDGEWORTH, J. D., DE LENCASTRE, H., PARKHILL, J., PEACOCK, S. J. & BENTLEY, S. D. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science,* 327**,** 469-74.

JOLLEY, K. A., BRAY, J. E. & MAIDEN, M. C. J. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res,* 3**,** 124.

LI, B. & WEBSTER, T. J. 2018. Bacteria antibiotic resistance: New challenges and opportunities for implant-associated orthopedic infections. *J Orthop Res,* 36**,** 22-32.

NASHEV, D., TOSHKOVA, K., SALASIA, S. I., HASSAN, A. A., LAMMLER, C. & ZSCHOCK, M. 2004. Distribution of virulence genes of Staphylococcus aureus isolated from stable nasal carriers. *FEMS Microbiol Lett,* 233**,** 45-52.

PEACOCK, S. J., MOORE, C. E., JUSTICE, A., KANTZANOU, M., STORY, L., MACKIE, K., O'NEILL, G. & DAY, N. P. J. 2002. Virulent Combinations of Adhesin and Toxin Genes in Natural Populations of Staphylococcus aureus. *Infection and Immunity,* 70**,** 4987-4996.

RASIGADE, J. P. & VANDENESCH, F. 2014. Staphylococcus aureus: a pathogen with still unresolved issues. *Infect Genet Evol,* 21**,** 510-4.

ROCA, I., AKOVA, M., BAQUERO, F., CARLET, J., CAVALERI, M., COENEN, S., COHEN, J., FINDLAY, D., GYSSENS, I., HEUER, O. E., KAHLMETER, G., KRUSE, H., LAXMINARAYAN, R., LIEBANA, E., LOPEZ-CERERO, L., MACGOWAN, A., MARTINS, M., RODRIGUEZ-BANO, J., ROLAIN, J. M., SEGOVIA, C., SIGAUQUE, B., TACCONELLI, E., WELLINGTON, E. & VILA, J. 2015. The global threat of antimicrobial resistance: science for intervention. *New Microbes New Infect,* 6**,** 22-9.

ROCHE, F. M., MEEHAN, M. & FOSTER, T. J. 2003. The Staphylococcus aureus surface protein SasG and its homologues promote bacterial adherence to human desquamated nasal epithelial cells. *Microbiology,* 149**,** 2759-67.

ROUARD, C., GARNIER, F., LERAUT, J., LEPAINTEUR, M., RAHAJAMANANAV, L., LANGUEPIN, J., PLOY, M.-C., BOURGEOIS-NICOLAOS, N. & DOUCETPOPULAIRE, F. 2018. Emergence and Within-Host Genetic Evolution of MethicillinResistant Staphylococcus aureus Resistant to Linezolid in a Cystic Fibrosis Patient. *Antimicrobial Agents and Chemotherapy,* 62**,** e00720-18.

RUIMY, R., ANGEBAULT, C., DJOSSOU, F., DUPONT, C., EPELBOIN, L., JARRAUD, S., LEFEVRE, L. A., BES, M., LIXANDRU, B. E., BERTINE, M., MINIAI, A. E., RENARD, M., BETTINGER, R. M., LESCAT, M., CLERMONT, O., PEROZ, G., LINA, G., TAVAKOL, M., VANDENESCH, F., VAN BELKUM, A., ROUSSET, F. & ANDREMONT, A. 2010. Are the Host Genetics Predominant Determinant of Persistent Nasal *Staphylococcus aureus* Carriage In Humans? *The Journal of Infectious Diseases,* 202**,** 924- 934.

SCHMIDT, A., BENARD, S. & CYR, S. 2015. Hospital Cost of Staphylococcal Infection after Cardiothoracic or Orthopedic Operations in France: A Retrospective Database Analysis. *Surg Infect (Larchmt),* 16**,** 428-35.

SHUKLA, S. K., KAROW, M. E., BRADY, J. M., STEMPER, M. E., KISLOW, J., MOORE, N., WROBLEWSKI, K., CHYOU, P. H., WARSHAUER, D. M., REED, K. D., LYNFIELD, R. & SCHWAN, W. R. 2010. Virulence genes and genotypic associations in nasal carriage, community-associated methicillin-susceptible and methicillin-resistant USA400 Staphylococcus aureus isolates. *J Clin Microbiol,* 48**,** 3582-92.

TAYLOR, T. A. & UNAKAL, C. G. 2018. *Staphylococcus Aureus*, StatPearls Publishing.

VAN WAMEL, W. J., ROOIJAKKERS, S. H., RUYKEN, M., VAN KESSEL, K. P. & VAN STRIJP, J. A. 2006. The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of Staphylococcus aureus are located on beta-hemolysin-converting bacteriophages. *J Bacteriol,* 188**,** 1310-5.

VERKAIK, N. J., BENARD, M., BOELENS, H. A., DE VOGEL, C. P., NOUWEN, J. L., VERBRUGH, H. A., MELLES, D. C., VAN BELKUM, A. & VAN WAMEL, W. J. 2011. Immune evasion cluster-positive bacteriophages are highly prevalent among human Staphylococcus aureus strains, but they are not essential in the first stages of nasal colonization. *Clin Microbiol Infect,* 17**,** 343-8.

WERTHEIM, H. F. L., MELLES, D. C., VOS, M. C., VAN LEEUWEN, W., VAN BELKUM, A., VERBRUGH, H. A. & NOUWEN, J. L. 2005. The role of nasal

carriage in Staphylococcus aureus infections. *The Lancet Infectious Diseases,* 5**,** 751-762.

WERTHEIM, H. F. L., WALSH, E., CHOUDHURRY, R., MELLES, D. C., BOELENS, H. A. M., MIAJLOVIC, H., VERBRUGH, H. A., FOSTER, T. & VAN BELKUM, A. 2008. Key Role for Clumping Factor B in Staphylococcus aureus Nasal Colonization of Humans. *PLoS Medicine* 5**,** e17.

WILLIAMS, I., ALFRED VENABLES, T. W., LLOYD, D., FRANK PAUL, I. F. & CRITCHLEY, I. 1997. The effects of adherence to silicone surfaces on antibiotic susceptibility in Staphylococcus aureus. *Microbiology,* 143**,** 2407-2413.

http://www.applied-maths.com/
https://pubmlst.org/saureus/
http://www.mgc.ac.cn/VFs/main.htm
http://www.genomicepidemiology.org/
https://github.com/tseemann/abricate

# Chapter 3

## Retrospective Definition of *Clostridioides difficile* PCR Ribotypes on the Basis of Whole Genome Polymorphisms: A Proof of Principle Study

Manisha Goyal[1], Lysiane Hauben[2], Hannes Pouseele[3], Magali Jaillard[4], Katrien De Bruyne[2], Alex van Belkum[1]* and Richard Goering[5]

[1] BioMérieux, Open Innovation and Partnerships, 3 Route du Port Michaud, 38390 La Balme Les Grottes, France

[2] BioMérieux, Applied Maths NV, 9830 Sint-Martens-Latem, Belgium

[3] BioMérieux, Industry, 69290 Craponne, France

[4] BioMérieux, 69280 Marcy l'Etoile, France

[5] Department of Medical Microbiology and Immunology, Creighton University School of Medicine, 2500 California Plaza, Omaha, NE 68178, USA

## Abstract

*Clostridioides difficile* is a cause of health care-associated infections. The epidemiological study of *C. difficile* infection (CDI) traditionally involves PCR ribotyping. However, ribotyping will be increasingly replaced by whole genome sequencing (WGS). This implies that WGS types need correlation with classical *C. difficile*) in order to perform retrospective clinical studies. Here, we selected genomes of hyper-virulent *C. difficile* strains of RT001, RT017, RT027, RT078, and RT106 to try and identify new discriminatory markers using in silico ribotyping PCR and De Bruijn graph-based Genome Wide Association Studies (DBGWAS). First, in silico ribotyping PCR was performed using reference primer sequences and 30 *C. difficile* genomes of the five different RTs identified above. Second, discriminatory genomic markers were sought with DBGWAS using a set of 160 independent *C. difficile* genomes (14 ribotypes). RT-specific genetic polymorphisms were annotated and validated for their specificity and sensitivity against a larger dataset of 2425 *C. difficile* genomes covering 132 different RTs. In silico PCR ribotyping was unsuccessful due to non-specific or missing theoretical RT PCR fragments. More successfully, DBGWAS discovered a total of 47 new markers (13 in RT017, 12 in RT078, 9 in RT106, 7 in RT027, and 6 in RT001) with minimum q-values of 0 to 7.40 × 10−5, indicating excellent marker selectivity. The specificity and sensitivity of individual markers ranged between 0.92 and 1.0 but increased to 1 by combining two markers, hence providing undisputed RT identification based on a single genome sequence. Markers were scattered throughout the *C. difficile* genome in intra- and intergenic regions. We propose here a set of new genomic polymorphisms that efficiently identify five hyper-virulent RTs utilizing WGS data only. Further studies need to show whether this initial proof-of-principle observation can be extended to all 600 existing RTs.

## Introduction

*Clostridioides difficile* (*C. difficile*), formerly known as *Clostridium difficile*, is an anaerobic, spore-forming Gram-positive bacterial species that can survive in harsh environments. It can withstand high temperatures, exposure to ultraviolet light, toxic chemicals, and exposure to antibiotics. Colonization by *C. difficile* is asymptomatic. The development of disease is mostly driven by host factors and disruption of the gut microbiome by frequent consumption of antibiotics (McFarland, 2009, Walk et al., 2012, Walker et al.,

2013). Toxigenic strains of *C. difficile* can be a lethal cause of *C. difficile* infection (CDI), which is commonly associated with post antibiotic diarrhea (Lessa et al., 2012, Wiegand et al., 2012). *C. difficile* is present in the environment and can be transmitted to patients or healthcare workers through contact with contaminated surfaces. Inter-human spread mainly occurs through the fecal–oral route. *C. difficile* spores are intrinsically resistant to antibiotics and remain viable during antibiotic treatment. Clindamycin, cephalosporins, and fluoroquinolones are considered as major antibiotics associated with CDI (Deshpande et al., 2013). Food or water contamination, gastric acid-suppression, and asymptomatic carriage in the community are the potential risk factors of community acquired CDI (Namiki and Kobayashi, 2018). One-third of the total CDI burden occurring in the USA in 2011 was community-associated (Lessa et al., 2015). CDI caused half a million hospital-acquired infections and 29,000 deaths in 2012 in the United States (Lessa et al., 2015) and approximately 40,000 cases of CDI in Europe (Davies et al., 2014). Only a limited number of studies reported emerging CDI in Asia (Borren et al., 2017). The increasing incidence of CDI and rapid evolution of antibiotic resistance in *C. difficile* has become a global threat to public health (Balsells et al., 2019, CDC, 2017, Mills et al., 2018).

CDI diagnosis allows early pathogen isolation and treatment of infection, thereby reducing the potential of CDI transmission. Various diagnostic procedures for CDI are available, including toxigenic culture, cell cytotoxic neutralization assay, glutamate dehydrogenase assay, the detection of toxins by enzyme immunoassays, nucleic acid amplification-based molecular tests, etc. (Burnham and Carroll, 2013, Eckert et al., 2013, Krutova et al., 2019, Planche and Wilcox, 2011, She et al., 2009, Shetty et al., 2011 ). Still, epidemiological *C. difficile* strain typing is necessary to identify outbreaks within a hospital or the wider community and facilitates understanding of the dissemination of infections. Ribotyping is a classical technique for *C. difficile* typing initially based on hybridization patterns of conserved ribosomal RNA probe sequences (Chatterjee and Raval, 2019, Dingle and MacCannell, 2015). Ribotype (RT) analysis has also been extremely important in the long-term surveillance of CDI (Krutova et al., 2018). While traditional typing methods such as restriction endonuclease analysis (REA) and pulsed-field gel electrophoresis (PFGE) were widely used in the past, PCR-based ribotyping is the current method of choice for *C. difficile* typing (Bidet et al., 1999, Collins et al., 2015). PCR ribotyping is dependent on the amplification of the intergenic spacer region (ISR) between 16S and 23S rRNA genes (Indra et al., 2010, Indra et al., 2008, Janezic, 2016, Waslawski et al., 2013). Since most bacterial species encode multiple ribosomal alleles in their genomes, multiple fragments of different

lengths are amplified when different species but also different strains are considered (Indra et al., 2010, Indra et al., 2008). There are still considerable constraints on PCR ribotyping including elevated costs, a higher probability of false-positive results, and a lack of interlaboratory portability (Borren et al., 2017, Martinez-Melendez et al., 2017, Polage et al., 2015).

Bacterial whole genome sequencing (WGS) has the potential to provide more detailed epidemiological information than classical PCR ribotyping (Collins et al., 2015, Janezic, 2016). To further explore a WGS-based approach to *C. difficile* typing, backward compatibility with PCR ribotyping is essential (Fawley et al., 2015). Previous studies reported the association of RT001, RT017, and RT027 with lethal CDI and considered those isolates as hyper-virulent (Arvand et al., 2009, Bauer et al., 2011). A survey conducted in the North East of England concluded that RT001, RT027, and RT106 were among the most prevalent and dangerous clones (Vanek et al., 2012). In the United States and Europe, RT001, RT014, RT020, RT027, and RT078 have been identified as predominant (Giancola et al., 2018 , Howell et al., 2010 ). RT017 is a globally emerging toxigenic RT and can be found on almost every continent (Imwattana et al., 2019, Kim et al., 2016). Thus, here, we tested both in silico PCR ribotyping and the De Bruijn graph-based Genome Wide Association Study (DBGWAS) (Jaillard et al., 2018a) for their capacity to perform retrospective PCR ribotyping for *C. difficile* RT001, RT017, RT027, RT078, and RT106. These strains were chosen as a test set representing global, long-term circulating and clinically relevant epidemic strains. The primary study goal was to develop a proof-of-principle procedure for sequence-based *C. difficile* strain typing with retrospective compatibility to established PCR RTs.

## Materials and Methods

### In Silico PCR-Based Ribotyping

We performed in silico PCR using canonical ribotyping PCR primers. Based on the reference sequences 16S-USA and 23S-USA (Table 3- 1), in silico PCR was performed using the subsequence search tool in BioNumerics v7.6 software (Applied Maths NV, Sint Martens-Latem, Belgium). Besides 7 genomes obtained from Creighton University, 23 *C. difficile* genomes of five selected RTs were downloaded from NCBI to verify in silico amplification of the ISR region (Supplementary Table S1).

***Table 3-1:*** *Primer pair used for in silico PCR-based ribotyping of Clostridioides difficile .*

| Primer | Gene Target | GenBank Accession No. | Sequence (5'–3') | Tm (°C) | Reference |
|---|---|---|---|---|---|
| 16S-USA (Forward) | 16S rRNA gene | FN545816 | (12293)GTGCGGCTGGATCACCTCCT (12312) | 71.0 | Xiao et al., 2012 (46) |
| 23S-USA (Reverse) | 23S rRNA gene | FN545816 | (12621)CCCTGCACCCTTAATAACTTGACC (12598) | 67.1 | |

## DBGWAS-Mediated Discovery New RT-Specific Markers

A total number of 160 *C. difficile* genome assemblies (training set) of 14 different RTs including hyper-virulent RT001, RT017, RT027, RT078, and RT106 were used for the discovery of unique RT genomic markers (Table 3- 2). This small training set allowed for the development of discriminatory markers to characterize the five major RTs among the 14 different RTs. Genomes were collected from the National Center for Biotechnology and Information (NCBI; www.ncbi.nlm.nih.gov), Creighton University, and the Enterobase databases (https://enterobase.warwick.ac.uk). Metadata of these genomes are summarized in Supplementary Table S2.

***Table 3-2:*** *C. difficile ribotypes included in the training dataset along with the number of genomes and their source of availability.*

| *C. difficile* Ribotype | Number of Genomes | Source |
|---|---|---|
| RT001 | 24 | Enterobase, NCBI, Creighton University |
| RT002 | 2 | NCBI, Creighton University |
| RT003 | 19 | NCBI, Creighton University |
| RT005 | 19 | NCBI, Creighton University |
| RT010 | 3 | NCBI, Creighton University |
| RT014 | 11 | NCBI, Creighton University |
| RT015 | 2 | NCBI, Creighton University |
| RT017 | 15 | NCBI, Creighton University |
| RT023 | 3 | NCBI, Creighton University |
| RT027 | 15 | NCBI, Creighton University |
| RT046 | 4 | NCBI, Creighton University |
| RT078 | 15 | NCBI, Creighton University |
| RT106 | 22 | Enterobase, NCBI, Creighton University |
| RT126 | 6 | NCBI, Creighton University |
| **TOTAL** | **160** | |

To identify associations between variant genetic loci and PCR RT, a hypothesis-free DBGWAS method was used. DBGWAS defines genetic variants linked to phenotypic traits via single nucleotide polymorphism (SNP), insertions, deletions, and consequences of recombination (Chewapreecha et al., 2014, Jaillard et al., 2018a). We used an open source tool (https://gitlab.com/leoisl/dbgwas) (Jaillard et al., 2018a) The tool is able to cover variants in coding as well as non-coding regions of bacterial genomes. DBGWAS was performed keeping the tool parameters in the default setting for different *C. difficile* ribotypes (RT001, RT017, RT027, RT078, and RT106) and their RT-specific genetic variants observed in the training set. Each *C. difficile* RT was considered independently in this study. DBGWAS identified short signature sequences called (overlapping) k-mers, yielding a compact summary of all variations across a set of genomes (Jaillard et al., 2018a). Overlapping k-mers are called unitigs and were selected on the basis of their specific and unique presence in a particular RT. Q-values define test sensitivity and specificity and are Benjamini–Hochberg-transformed p-values for controlling the false-positive results in case of multiple testing (Benjamini et al., 1995, Jaillard et al., 2018a). R scripting was used to deal with large matrices defining the presence (1) or absence (0) of extracted unitigs in the training set of *C. difficile* genomes.

**Validation of Markers**

Validation of novel unitig markers was performed by means of BLAST searches against the test set of genomes (Table 3- 3). A wide range of 2425 genomes covering 132 different *C. difficile* RTs was downloaded (Frentrup et al., 2019) and processed using a Linux shell script. These sequences represented PCR ribotyped strains from different countries and clinical and environmental specimens for which phylogenetic analyses were already performed by Frentrup et al. (Frentrup et al., 2019) (Supplementary Table S3). A database of this test set was created to perform local command line BLAST searches against the set of significant unitigs identified above. The specificity of all the unitigs was tested using strict parametric filters of 100% coverage and identity.

***Table 3-3:*** *C. difficile ribotypes downloaded from the Enterobase database as a test dataset and the number of genomes included in each ribotype.*

| Ribotype | Count | Ribotype | Count | Ribotype | Count | Ribotype | Count |
|---|---|---|---|---|---|---|---|

| Ribotype | Count | Ribotype | Count | Ribotype | Count | Ribotype | Count |
|---|---|---|---|---|---|---|---|
| RT001 | 206 | RT046 | 3 | RT127 | 1 | RT375 | 1 |
| RT002 | 53 | RT049 | 9 | RT129 | 1 | RT404 | 15 |
| RT003 | 11 | RT050 | 4 | RT137 | 1 | RT413 | 13 |
| RT005 | 14 | RT051 | 1 | RT138 | 1 | RT446 | 2 |
| RT006 | 1 | RT053 | 5 | RT149 | 1 | RT449 | 2 |
| RT009 | 2 | RT054 | 2 | RT150 | 1 | RT451 | 1 |
| RT010 | 7 | RT056 | 5 | RT153 | 1 | RT453 | 1 |
| RT011 | 3 | RT058 | 1 | RT156 | 1 | RT454 | 1 |
| RT012 | 45 | RT060 | 1 | RT157 | 1 | RT456 | 1 |
| RT013 | 1 | RT062 | 2 | RT158 | 1 | RT470 | 1 |
| RT014 | 113 | RT063 | 1 | RT176 | 13 | RT500 | 21 |
| RT015 | 36 | RT066 | 3 | RT193 | 1 | RT534 | 1 |
| RT017 | 272 | RT067 | 1 | RT194 | 1 | RT547 | 1 |
| RT018 | 55 | RT069 | 1 | RT212 | 1 | RT559 | 1 |
| RT019 | 1 | RT070 | 4 | RT220 | 4 | RT563 | 1 |
| RT020 | 44 | RT072 | 1 | RT225 | 1 | RT569 | 1 |
| RT022 | 1 | RT073 | 2 | RT226 | 1 | RT581 | 1 |
| RT023 | 16 | RT075 | 1 | RT236 | 3 | RT585 | 1 |
| RT024 | 1 | RT076 | 2 | RT238 | 1 | RT586 | 1 |
| RT026 | 6 | RT077 | 1 | RT239 | 2 | RT591 | 1 |
| RT027 | 652 | RT078 | 492 | RT241 | 5 | RT598 | 8 |
| RT029 | 3 | RT081 | 2 | RT244 | 9 | RT614 | 1 |
| RT031 | 2 | RT083 | 1 | RT251 | 1 | RT620 | 2 |
| RT032 | 1 | RT084 | 2 | RT262 | 1 | RT629 | 1 |
| RT033 | 5 | RT087 | 8 | RT284 | 1 | RT651 | 1 |
| RT035 | 2 | RT090 | 1 | RT289 | 1 | RT666 | 1 |
| RT036 | 1 | RT094 | 1 | RT290 | 1 | RT668 | 1 |
| RT037 | 1 | RT102 | 1 | RT305 | 1 | RT678 | 1 |
| RT039 | 5 | RT103 | 2 | RT316 | 1 | RT708 | 1 |
| RT042 | 2 | RT106 | 55 | RT321 | 1 | RT719 | 1 |
| RT043 | 2 | RT117 | 2 | RT328 | 2 | RT720 | 1 |
| RT044 | 2 | RT125 | 1 | RT336 | 1 | RT721 | 1 |
| RT045 | 2 | RT126 | 79 | RT356 | 8 | RT722 | 1 |

## Statistically Reliable Ribotype Prediction

To evaluate the potential typing significance of the unitigs as compared to the classical ribotyping of *C. difficile* strains, sensitivity and specificity (selectivity) were computed for all the unitigs (Baratloo et al., 2015). The efficiency of GWAS can be measured by assessing the false discovery rate (FDR) (Bradbury et al., 2011). To increase the potential typing significance of our new method, combination statistics were performed. Sensitivity and specificity were also computed for certain combinations of two or more selected unitig sequences. The parameters defined were, next to the FDR, TP (true-positives, correctly predicting positive values, e.g., number of true RT017 predicted as RT017), FP (false-positives, missed negative values, e.g., number of non-RT017 genomes still predicted as RT017), FN (false-negatives, missed positive values), and TN (true-negatives, correctly rejected values).

**Functional Annotation of Unitigs**

Selected unitigs were annotated using BLASTn alignment. Well-characterized *C. difficile* genomes were used as a reference to locate these new markers. Specific annotation for each marker was filtered out using minimum E-value, 100% identity, 100% coverage, and 0 gap score.

**Results and Discussion**

**In Silico PCR**

The visualization of amplified DNA sequences from the intergenic region between 16S and 23S ribosomal genes is the current Gold Standard for *C. difficile* typing (Indra et al., 2010, Xiao et al., 2012). In our study, in silico PCR for 30 randomly selected, well-characterized *C. difficile* genome sequences was essentially unsuccessful (Figure 3- 1 and Supplementary Table S1). Genome sequences included generated insufficient numbers of differently sized fragments. The fragment sizes that were calculated were verified with the online tool available at http://insilico.ehu.es/PCR (Borren et al., 2017). On the other hand, more recently sequenced *C. difficile* genomes were showing only one or even none of the expected amplified fragments (Supplementary Table S1). There is a substantial possibility that the PCR ribotyping fragments observed upon laboratory experimentation for these strains may not derive from ISR variants but rather from random amplification. Thus, in silico PCR failed to generate reliable RTs which prompted us to explore the feasibility of DBGWAS-

based typing. Of note, we presume here that NGS-based methods are very likely to be more reliable than any of the many other molecular typing methods.

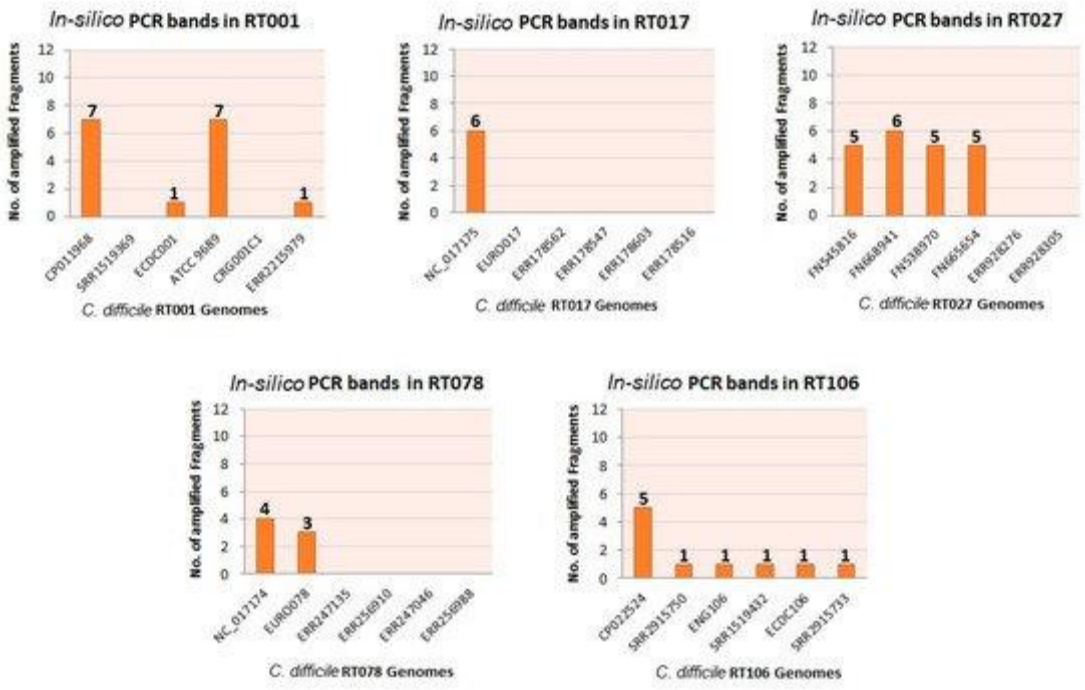
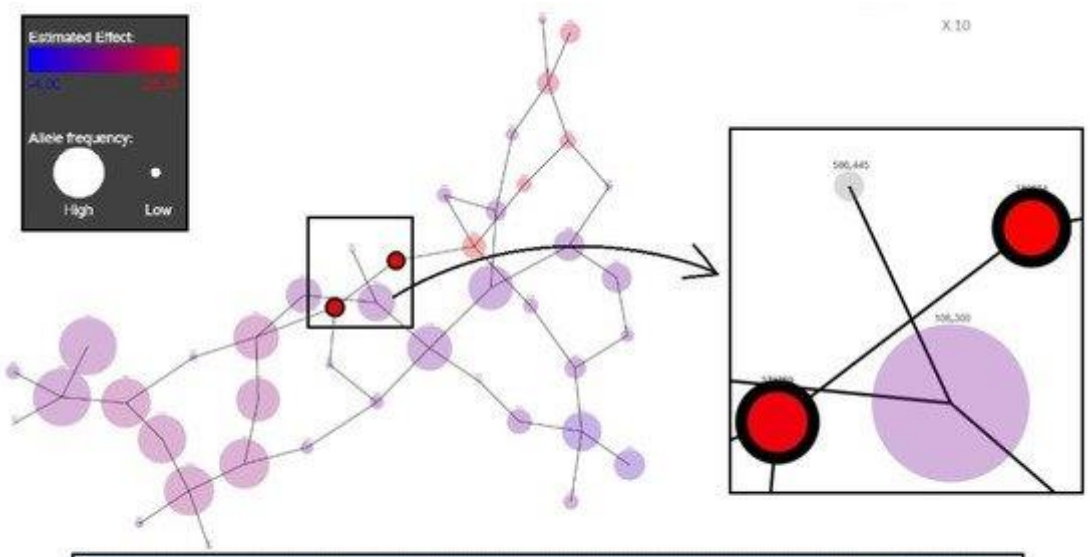



***Figure 3-1:*** *In silico ribotyping of different C. difficile genome sequences using the ISR 16S and 23S USA primer pair. The five panels represent the results obtained for examples of five different ribotypes. Bar graphs show the number of theoretical PCR bands (vertical axis, number of bands labeled on each bar) in the ribosomal region of respective genome sequences (horizontal axis), whereas the genomes without any fragments depict the complete absence of primer binding sites in those genomes. Note that the expected outcome would be an identical number of fragments for each of the strains belonging to a single ribotype. We indicated this number as the first marker of reproducibility; it has to be stated that besides this variation of numbers of fragments, the size of the fragment was also determined as a variable as well.*

**New Genotyping Markers**

A total number of 47 RT-specific unitigs (13 for RT017, 12 for RT078, 9 for RT106, 7 for RT027, and 6 for RT001) were identified. The unitigs shared an average length of 56 base pairs (Table 3- 4). DBGWAS generated compacted De Bruijn graphs (cDBG) containing the specific unitigs as nodes defining a genotypic association between a particular RT and the *C. difficile* genomes included (Figure 3- 2). Unitigs that were specific for a particular RT were color-coded according to their association to the RT (red for positive association, blue for negative association) and minimum q-values were provided by

subgraphs. Q-values for selected unitigs ranged from 0 to 7.40 × 10−5. Significantly associated unitigs were extracted as FASTA formatted sequences (Supplementary Table S4).




| Node ID | Allele freq | Pheno0 (<=0) | Pheno1 (>0) | NA | Annotation | Significant? | q-Value | Est effect |
|---|---|---|---|---|---|---|---|---|
| n180654 | 23 | 0/126 | 23/24 | 0/0 | | Yes | 8.82 x $10^{-39}$ | 9.98 x $10^{-01}$ |
| n373759 | 26 | 3/126 | 23/24 | 0/0 | | Yes | 2.96 x $10^{-28}$ | 8.93 x $10^{-01}$ |

***Figure 3-2:*** *Compacted De Bruijn graphs (cDBG) generated by De Bruijn graph-based Genome Wide Association Studies (DBGWAS) for RT001 genome sequences. The figure illustrates the significance of the nodes (representing the selective sequences called unitigs), which are denoted by their estimated effect ranging from high (28.304; red) to low (4.00; blue). Allele frequency is represented by the size of the node. The table explains that from the two selected significant nodes in terms of their association with ribotype, the node on the top right (n180654) is specific to RT001 (called Pheno 1 in the table) and completely absent in the other ribotypes in the training set (Pheno 0). Additionally, the q-value linked to the first node is very significantly below 0.05 and hence, the estimated effect is high (represented by the red color of the node).*

***Table 3-4:*** *Number of unique markers identified for each C. difficile ribotype, their average length, and annotation.*

| Ribotype | Number of Markers | Average Length (Base Pairs) | Annotation (Number of Unitigs) |
|---|---|---|---|
| RT001 | 06 | 59 | 1. Intergenic (4)<br>2. tRNA uridine-5-carboxymethylaminomethyl synthesis enzyme MnmG (1)<br>3. rRNA-23S ribosomal RNA (1 excluded from the list)<br>4. Unknown (1) |
| RT017 | 13 | 69 | 1. Intergenic (3)<br>2. Membrane spanning protein (1)<br>3. Ribosome small subunit-dependent GTPase A (1)<br>4. Hypothetical protein (1)<br>5. EAL domain-containing protein (2)<br>6. Glutamate 2,3-aminomutase (1)<br>7. MurR/RpiR family transcriptional regulator (1)<br>8. GGDEF domain-containing protein (1) |

| Ribotype | Number of Markers | Average Length (Base Pairs) | Annotation (Number of Unitigs) |
|---|---|---|---|
| | | | 9. Glyoxalase-like domain protein (1)<br>10. Radical SAM protein (1) |
| RT027 | 07 | 53 | 1. Collagen-like exosporium glycoprotein BclA2 (1)<br>2. Intergenic (5)<br>3. Unknown (1) |
| RT078 | 12 | 42 | 1. IS200/IS605 family element transposase accessory protein TnpB (1)<br>2. Spore surface glycoprotein BclB (4)<br>3. Collagen-like exosporium glycoprotein (BclA2) (3)<br>4. Unknown (partial with ABC transporter permease) (1)<br>5. Intergenic (1)<br>6. Site-specific integrase (1)<br>7. S8 family peptidase (1) |
| RT106 | 09 | 55 | 1. Intergenic (3)<br>2. ABC transporter permease (2)<br>3. Hypothetical Protein (1)<br>4. 3-Hydroxybutyryl-CoA dehydrogenase (1)<br>5. Potassium transporter (2) |

## Validation of Markers

Unitigs showing 100% identity in all genomes belonging to a single RT in the validation set demonstrated the efficiency of these unique patterns to carry out in silico ribotyping. Although the individual unitig-based characterization of *C. difficile* strains was not absolute, it allowed RT determination with approximate sensitivity and specificity of between 0.90 and 1.0 (Figure 3- 3). FDR for all the unitigs for RT017 was the lowest (0.06) followed by RT027 (0.08), RT078 (0.23), RT001 (0.30), and RT106 (0.46) (Figure 3- 3).

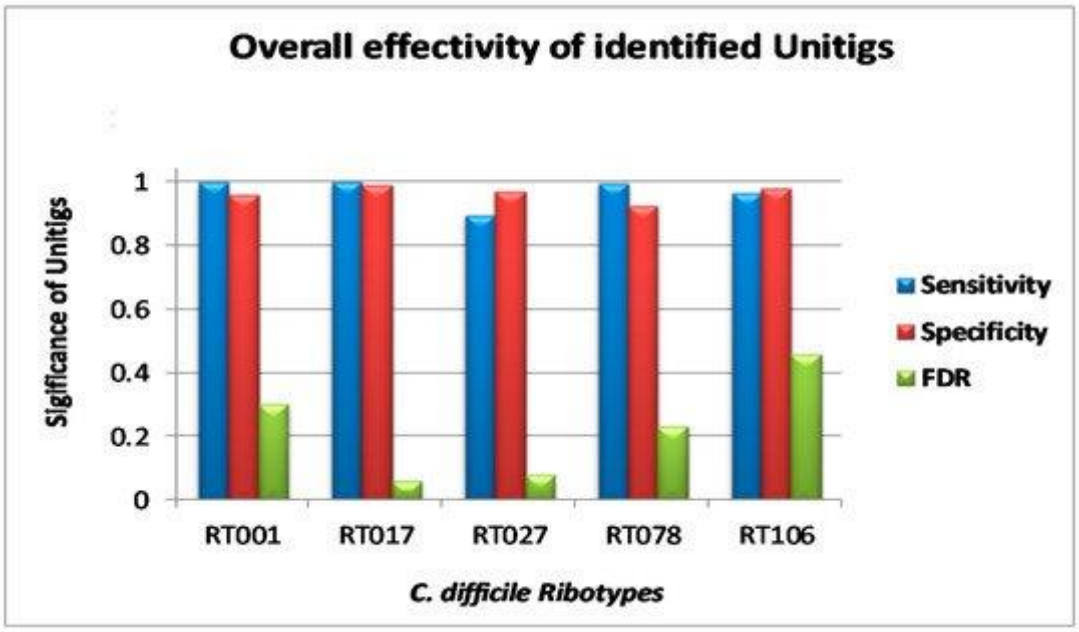



***Figure 3-3:*** *Statistical comparison of genome typing efficiency of discovered unique patterns.*

Some of the unitigs were shared by closely related RTs. Unitigs for RT001 were able to identify the genetically closely related RT087, RT241, and RT012, which altogether form a clonal complex (CC) 141 (Frentrup et al., 2019). One of the markers identified for RT017 showed no false-positives or false negatives. Other markers for RT017 initially generated a small number of false-positives, but 100% true-positives in the validation dataset. Markers for RT078 identified 78 out of 79 isolates of RT126 and all of the RT413 strains from the test dataset, likely due to the close genetic relatedness of these RTs (CC 1) (Alvarez-Perez et al., 2017, Frentrup et al., 2019, Schneeberg et al., 2013). Unitig sequences for RT106 were also able to identify *C. difficile* RT500 along with RT106 from the test set. Phylogenetic grouping of *C. difficile* genomes (Frentrup et al., 2019) showed that *C. difficile* core genome multi locus sequence typing (cgMLST) of RT106 and RT500 (CC 22) generated completely indistinguishable groupings. Considering closely related strains as true-positives based on their respective RTs, the FDRs for each unitig subset were found to be smaller, underscoring the biological consistency of the results. Adding genomes of RT413, RT126, and RT500 to the training set resulted in a decreased FDR rate. The continuously increasing number of publicly available *C. difficile* genome sequences will provide substantial opportunities for improvement of our new characterization technique.

**Marker Combination Study**

For the ribotypes RT027, RT078, RT106, and RT001, every possible combination of RT-specific unitigs was created and tested for statistical significance. A combination of two unitigs was shown to increase sensitivity and specificity up to 1 and to reduce the FDR to 0.05 (Figure 3- 4A–D). Each combination was defined on the basis of logical operators "AND/OR". The AND operator symbolizes that both the markers in a combination need to be present with 100% identity, whereas the OR operator means that any one of the two markers in a combination need to be present at one time, again with 100% sequence identity. There is no combination required in the case of RT017. Conclusively, as clearly exemplified in Figure 3- 4A–D, in certain cases, the combination of markers improves RT testing by suppression of the false discovery rate. Marker's SEQ ID numbers and their sequences are given in Supplementary Table S4.

A

| RT027:Combination SEQ ID°1 OR SEQ ID°2 | |
|---|---|
| False Positive | 30 |
| True Positive | 578 |
| False Negative | 74 |
| True Negative | 1743 |
| Sensitivity | 0.89 |
| Specificity | 0.98 |
| False Discovery Rate | 0.05 |

B

| RT078: Combination SEQ ID°21 OR SEQ ID°24 | |
|---|---|
| False Positive | 41 |
| True Positive | 472 |
| False Negative | 20 |
| True Negative | 1892 |
| Sensitivity | 0.96 |
| Specificity | 0.98 |
| False Discovery Rate | 0.08 |

C

| RT106: Combination SEQ ID°40 AND SEQ ID°41 | |
|---|---|
| False Positive | 20 |
| True Positive | 52 |
| False Negative | 3 |
| True Negative | 2350 |
| Sensitivity | 0.95 |
| Specificity | 0.99 |
| False Discovery Rate | 0.28 |

D

| RT001: Combination SEQ ID°44 AND SEQ ID°47 | |
|---|---|
| False Positive | 11 |
| True Positive | 205 |
| False Negative | 1 |
| True Negative | 2208 |
| Sensitivity | 1.00 |
| Specificity | 1.00 |
| False Discovery Rate | 0.05 |

***Figure 3-4:*** *(A–D) Statistical reliability in terms of sensitivity, specificity, and false discovery rate (FDR) for the combination of two selected markers using OR operator for the identification of C. difficile RT027 (Panel A) and RT078 (Panel B). Panels C and D display similar analyses but then using the AND operator for identification of RT106 and RT001, respectively.*

**Functional Annotation of Markers**

Functional characterization of the regions from which our unitigs originated demonstrated that 34% of the unitigs were localized in intergenic regions (five for RT027, four for RT001, three for of RT017 and RT106 each, and one for RT078 (Figure 3- 5, Figure

3- 6, Figure 3- 7, Figure 3- 8 and Figure 3- 9). Six percent of all markers were left unannotated in RT001, RT027, and RT078 (one marker for each) (Table 3- 4). Only RT001 was identified with a unitig marker residing within the rRNA-23S ribosomal gene showing at least some correspondence with ribotyping (Figure 3- 5). This marker did not show sufficient diagnostic power and was thus not selected in the final set of markers. All other markers were observed to be scattered throughout the *C. difficile* genome. In RT078, one of these markers was identified in a mobile genetic element (Figure 3- 8). Mostly, genes and intergenic regions, apart from the conserved ribosomal ISR, were observed to play a potential role in the unitig-mediated *C. difficile* typing.

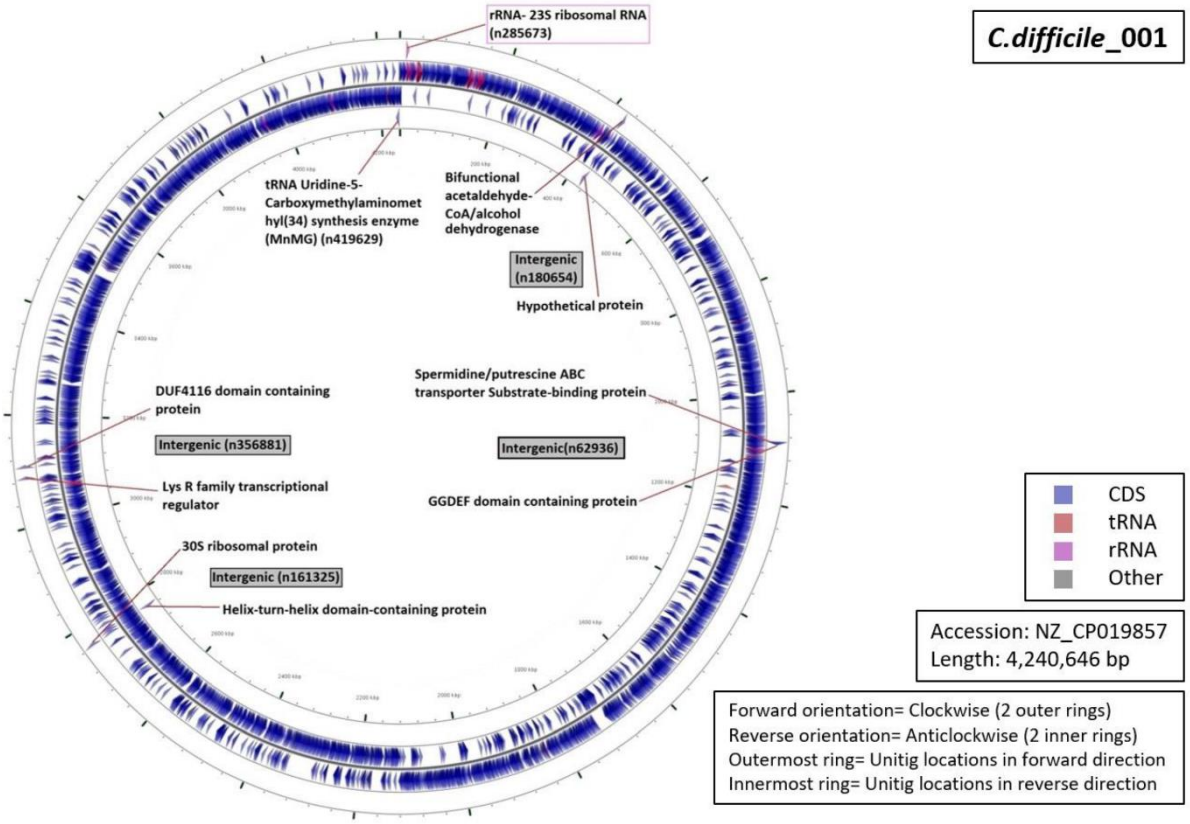



***Figure 3-5:*** *Functional annotation and location of DBGWAS markers on the reference genome of C. difficile RT001. Functional annotation and location of DBGWAS markers on the reference genome of C. difficile RT001. Both central rings represent the genome annotation (reverse inside, forward outside), while the outer and inner rings represent the signature sequences (unitigs) (reverse inside, forward outside).*

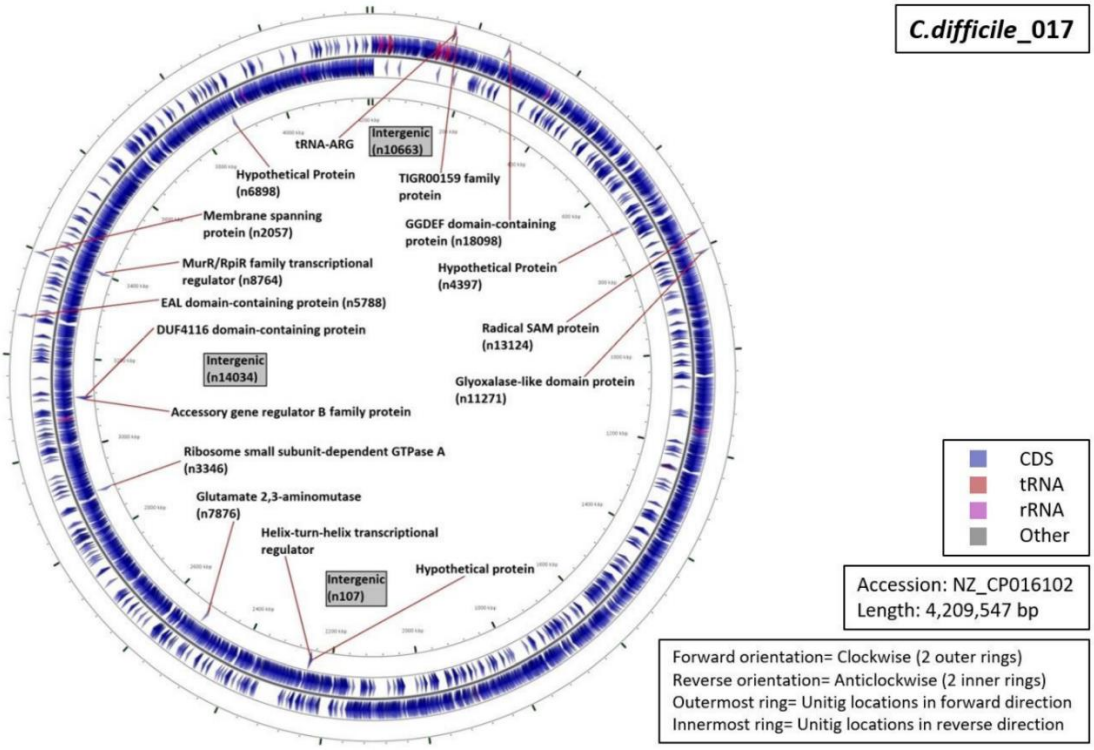



***Figure 3-6:*** *Functional annotation and location of DBGWAS markers on the reference genome of C. difficile RT017.*

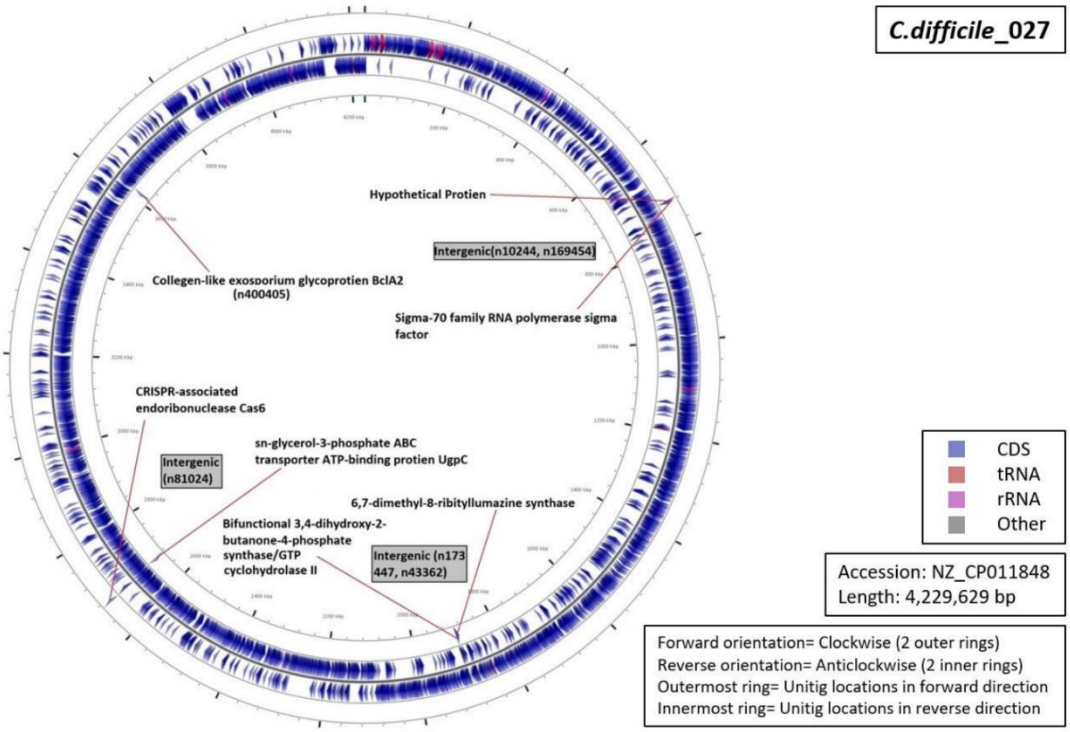



***Figure 3-7:*** *Functional annotation and location of DBGWAS markers on the reference genome of C. difficile RT027.*

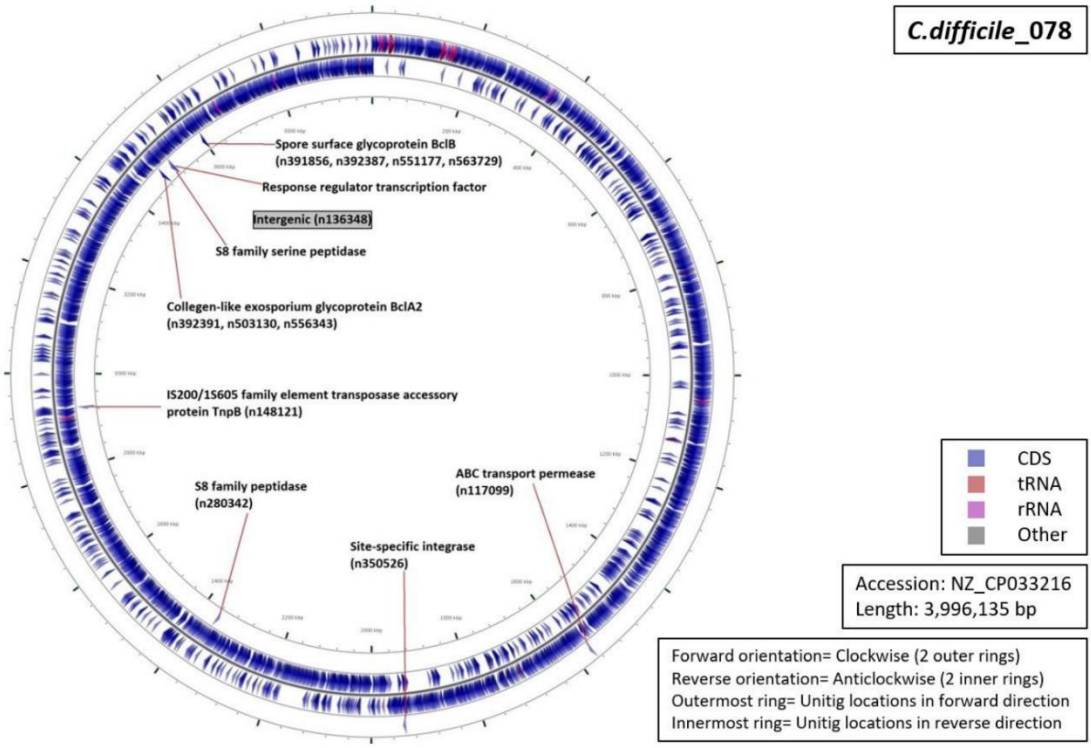



***Figure 3-8:*** *Functional annotation and location of DBGWAS markers on the reference genome of C. difficile RT078.*

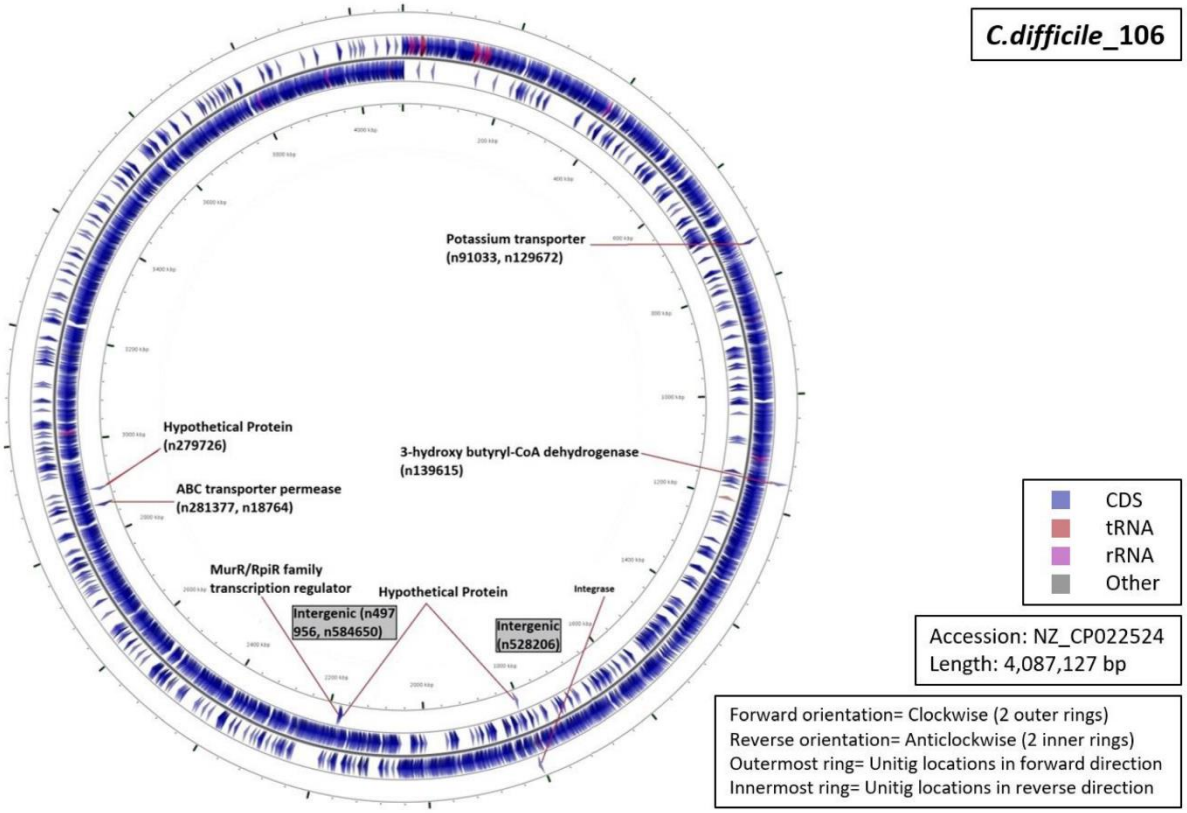



***Figure 3-9:*** *Functional annotation and location of DBGWAS markers on the reference genome of C. difficile RT106.*

## Conclusions

Strain typing has a proven value in monitoring the persistence and spread of bacterial pathogens in human populations. For *C. difficile*, PCR ribotyping is the current first choice but may be challenged now that genome sequencing is an option. No single-step test or algorithm is available so far for correlating *C. difficile* RTs with WGS data. This implies that there may be an issue with the correlation between WGS-based epidemiological analysis and PCR ribotyping for *C. difficile*. Here, we show that DBGWAS identified unique genomic markers that would suit that specific purpose. A combination of two unitigs led to 100% sensitive and specific discrimination between five important RTs. We believe that this approach is highly promising, providing a clear opportunity to define backward compatibility between classical RTs and WGS data.

## Supplementary Materials

The following are available online at https://www.mdpi.com/2075-4418/10/12/1078/s1. Table S1 describes a collection of genome sequences that were used for the in silico search of ribotypes based on amplification of tentative ISRs. Table S2 contains a training set of *C. difficile* genomes used for the initial DBGWAS. Table S3 contains the *C. difficile* genomes used for DBGWAS validation. Table S4 shows a review of all unitigs that are statistically significantly associated with specific ribotypes.

## Author Contributions

Conceptualization, A.v.B., M.G., L.H., M.J., K.D.B.and R.V.G.; Formal analysis, M.G.; Investigation, M.G., A.v.B., L.H., M.J. and R.V.G.; Methodology, M.G. and K.D.B.; Software, M.J.; Supervision, A.v.B.; Writing—original draft, M.G.; Writing—review & editing, A.v.B., L.H., H.P., M.J., K.D.B.and R.G. All authors have read and agreed to the published version of the manuscript.

## Funding

This research was supported and funded by bioMérieux, France; Creighton University School of Medicine, NE, USA; and the European Union's Horizon 2020 research and

innovation program entitled as Viral and Bacterial Adhesin Network Training (ViBrANT) under Marie Skłodowska-Curie Grant Agreement No. 765042.

## Acknowledgments

During this study, M.G., L.H., H.P., M.J., K.D.B. and A.v.B. were employees of bioMérieux, a company designing, developing, and marketing tests in the domain of infectious diseases. The company was not involved in the design of the current study and the opinions expressed are those of the authors and may be different from formal company opinions and policies. We thank our colleagues from bioMérieux and Creighton University who provided expertise and insight that greatly assisted the research.

## Conflicts of Interest

This research was conducted in association with RG from Creighton University School of Medicine, NE, USA which could be constructed as a potential conflict of interest.

## References

ALVAREZ-PEREZ, S., BLANCO, J. L., HARMANUS, C., KUIJPER, E. & GARCIA, M. E. 2017. Subtyping and antimicrobial susceptibility of Clostridium difficile PCR ribotype 078/126 isolates of human and animal origin. *Vet. Microbiol.***,** 15-22.

ARVAND, M., HAURI, A. M., ZAISS, N. H., WITTE, W. & BETTGE-WELLER, G. 2009. Clostridium difficile ribotypes 001, 017, and 027 are associated with lethal C. difficile infection in Hesse. *Euro Surveill.,* 14.

BALSELLS, E., SHI, T., LEESE, C., LYELL, I., BURROWS, J., WIUFF, C., CAMPBELL, H., KYAW, M. H. & NAIR, H. 2019. Global burden of Clostridium difficile infections: a systematic review and meta-analysis. *J Glob Health,* 9**,** 010407.

BARATLOO, A., HOSSEINI, M., NEGIDA, A. & G., E. A. 2015. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emerg (Tehran),* 3**,** 48-49.

BAUER, M. P., NOTERMANS, D. W., VAN BENTHEM, B. H., BRAZIER, J. S., WILCOX, M. H., RUPNIK, M., MONNET, D. L., VAN DISSEL, J. T., KUIJPER, E. J. & GROUP, E. S. 2011. Clostridium difficile infection in Europe: a hospital-based survey *Lancet Infect Dis.,* 377**,** 63–73.

BENJAMINI, Y., HOCHBERG, Y. & 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)***,** 289–300.

BIDET, P., BARBUT, F., LALANDE, V., BURGHOFFER, B. & PETIT, J.-C. 1999. Development of a new PCR-ribotyping method for Clostridium difficile based on ribosomal RNA gene sequencing. *FEMS microbiology letters,* 175**,** 261-266.

BORREN, N. Z., GHADERMARZI, S., HUTFLESS, S. & ANANTHAKRISHNAN, A. N. 2017. The emergence of Clostridium difficile infection in Asia: A systematic review and meta-analysis of incidence and impact. *PLoS One,* 12**,** e0176797.

BRADBURY, P., PARKER, T., HAMBLIN, M. T. & JANNINK, J. 2011. Assessment of Power and False Discovery Rate in Genome-Wide Association Studies using the BarleyCAP Germplasm. *Crop Sci.,* 51**,** 52-59.

BURNHAM, C. A. & CARROLL, K. C. 2013. Diagnosis of Clostridium difficile infection: an ongoing conundrum for clinicians and for clinical laboratories. *Clin Microbiol Rev,* 26**,** 604-30.

CDC 2017. Biggest Threats Antibiotic/Antimicrobial Resistance. *CDC*.

CHATTERJEE, S. & RAVAL, I. H. 2019. Chapter 32 - Pathogenic Microbial Genetic Diversity with Reference to Health,. *In:* SURAJIT DAS, H. R. D. (ed.) *Microbial Diversity in the Genomic Era.* Academic Press.

CHEWAPREECHA, C., MARTTINEN, P., CROUCHER, N. J., SALTER, S. J., HARRIS, S. R., MATHER, A. E., HANAGE, W. P., GOLDBLATT, D., NOSTEN, F. H., TURNER, C., TURNER, P., BENTLEY, S. D. & PARKHILL, J. 2014. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet,* 10**,** e1004547.

COLLINS, D. A., ELLIOTT, B. & RILEY, T. V. 2015. Molecular methods for detecting and typing of Clostridium difficile. *Pathology,* 47**,** 211-8.

DAVIES, K. A., LONGSHAW, C. M., DAVIS, G. L., BOUZA, E., BARBUT, F. & BARNA, Z. E. A. 2014. Underdiagnosis of Clostridium difficile across Europe: the European, multicentre, prospective, biannual, point-prevalence study of Clostridium difficile infection in hospitalised patients with diarrhoea (EUCLID). *Lancet Infect Dis.,* 14**,** 1208–19.

DESHPANDE, A., PASUPULETI, V., THOTA, P., PANT, C., ROLSTON, D. D., SFERRA, T. J., HERNANDEZ, A. V. & DONSKEY, C. J. 2013. Community-associated Clostridium difficile infection and antibiotics: a meta-analysis. *J Antimicrob Chemother,* 68**,** 1951-61.

DINGLE, T. C. & MACCANNELL, D. R. 2015. Chapter 9 - Molecular Strain Typing and Characterisation of Toxigenic Clostridium difficile. *Methods in Microbiology.*

ECKERT, C., JONES, G. & BARBUT, F. 2013. Diagnosis of Clostridium difficile infection: the molecular approach. *Future Microbiol,* 8**,** 1587-98.

FAWLEY, W. N., KNETSCH, C. W., MACCANNELL, D. R., HARMANUS, C., DU, T., MULVEY, M. R., PAULICK, A., ANDERSON, L., KUIJPER, E. J. & WILCOX, M. H. 2015. Development and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis PCR-ribotyping protocol for Clostridium difficile. *PLoS One,* 10**,** e0118150.

FRENTRUP, M., ZHOU, Z., STEGLICH, M., MEIER-KOLTHOFF, J. P., GÖKER, M., RIEDEL, T., BUNK, B., SPRÖER, C., OVERMANN, J., BLASCHITZ, M., INDRA, A., VON MÜLLER, L., KOHL, T. A., NIEMANN, S., SEYBOLDT, C., KLAWONN, F., KUMAR, N., LAWLEY, T. D., GARCÍA-FERNÁNDEZ, S., CANTÓN, R., DEL CAMPO, R., ZIMMERMANN, O., GROß, U., ACHTMAN, M. & NÜBEL, U. 2019.

GIANCOLA, S. E., WILLIAMS, R. J. & GENTRY, C. A. 2018 Prevalence of the Clostridium difficile BI/NAP1/027 strain across the United States Veterans Health Administration. . *Clin Microbiol Infect. ,* 24**,** 877-881.

HOWELL, M. D., NOVACK, V., GRGURICH, P., SOULLIARD, D., NOVACK, L., PENCINA, M. & TALMOR, D. 2010 Latrogenic gastric acid suppression and the risk of nosocomial Clostridium difficile infection. *Arch Intern Med. ,* 170**,** 784-90.

IMWATTANA, K., KNIGHT, D. R., KULLIN, B., COLLINS, D. A., PUTSATHIT, P., KIRATISIN, P. & RILEY, T. V. 2019. Clostridium difficile ribotype 017 -

characterization, evolution and epidemiology of the dominant strain in Asia. *Emerg Microbes Infect,* 8**,** 796-807.

INDRA, A., BLASCHITZ, M., KERNBICHLER, S., REISCHL, U., WEWALKA, G. & ALLERBERGER, F. 2010. Mechanisms behind variation in the Clostridium difficile 16S-23S rRNA intergenic spacer region. *J Med Microbiol,* 59**,** 1317-23.

INDRA, A., HUHULESCU, S., SCHNEEWEIS, M., HASENBERGER, P., KERNBICHLER, S., FIEDLER, A., WEWALKA, G., ALLERBERGER, F. & KUIJPER, E. J. 2008. Characterization of Clostridium difficile isolates using capillary gel electrophoresis-based PCR ribotyping. *J Med Microbiol,* 57**,** 1377-82.

JAILLARD, M., LIMA, L., TOURNOUD, M., MAHE, P., VAN BELKUM, A., LACROIX, V. & JACOB, L. 2018a. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet,* 14**,** e1007758.

JANEZIC, S. 2016. Direct PCR-Ribotyping of Clostridium difficile. *In:* ROBERTS, A. & MULLANY, P. (eds.) *Clostridium difficile. Methods in Molecular Biology.* New York, NY: Humana Press.

KIM, J., KIM, Y. & PAI, H. 2016. Clinical Characteristics and Treatment Outcomes of Clostridium difficile Infections by PCR Ribotype 017 and 018 Strains. *PLoS One,* 11**,** e0168849.

KRUTOVA, M., KINROSS, P., BARBUT, F., HAJDU, A., WILCOX, M. H., KUIJPER, E. J. & SURVEY, C. 2018. How to: Surveillance of Clostridium difficile infections. *Clin Microbiol Infect,* 24**,** 469-475.

KRUTOVA, M., WILCOX, M. H. & KUIJPER, E. J. 2019. A two-step approach for the investigation of a Clostridium difficile outbreak by molecular methods. *Clin Microbiol Infect.*

LESSA, F. C., GOULD, C. V. & MCDONALD, L. C. 2012. Current status of Clostridium difficile infection epidemiology. *Clin Infect Dis,* 55 Suppl 2**,** S65-70.

LESSA, F. C., MU, Y., BAMBERG, W. M., BELDAVS, Z. G., DUMYATI, G. K., DUNN, J. R., FARLEY, M. M., HOLZBAUER, S. M., MEEK, J. I., PHIPPS, E. C., WILSON, L. E., WINSTON, L. G., COHEN, J. A., LIMBAGO, B. M., FRIDKIN, S. K., GERDING, D. N. & MCDONALD, L. C. 2015. Burden of Clostridium difficile infection in the United States. *N Engl J Med,* 372**,** 825-34.

MARTINEZ-MELENDEZ, A., CAMACHO-ORTIZ, A., MORFIN-OTERO, R., MALDONADO-GARZA, H. J., VILLARREAL-TREVINO, L. & GARZA-GONZALEZ, E. 2017. Current knowledge on the laboratory diagnosis of Clostridium difficile infection. *World J Gastroenterol,* 23**,** 1552-1567.

MCFARLAND, L. V. 2009. Renewed interest in a difficult disease: Clostridium difficile infections—epidemiology and current treatment strategies. *Curr Opin Gastroenterol,* 25**,** 24–35.

MILLS, J. P., RAO, K. & YOUNG, V. B. 2018. Probiotics for prevention of Clostridium difficile infection. *Curr Opin Gastroenterol,* 34**,** 3–10.

NAMIKI, H. & KOBAYASHI, T. 2018. Long-term, low-dose of clarithromycin as a cause of community-acquired Clostridium difficile infection in a 5-year-old boy. *Oxf Med Case Reports,* 2018**,** omx106.

PLANCHE, T. & WILCOX, M. 2011. Reference assays for Clostridium difficile infection: one or two gold standards? *J Clin Pathol,* 64**,** 1-5.

POLAGE, C. R., GYORKE, C. E., KENNEDY, M. A., LESLIE, J. L., CHIN, D. L., WANG, S., NGUYEN, H. H., HUANG, B., TANG, Y. W., LEE, L. W., KIM, K., TAYLOR, S., ROMANO, P. S., PANACEK, E. A., GOODELL, P. B., SOLNICK, J. V. &

COHEN, S. H. 2015. Overdiagnosis of Clostridium difficile Infection in the Molecular Test Era. *JAMA Intern Med,* 175**,** 1792-801.

SCHNEEBERG, A., NEUBAUER, H., SCHMOOCK, G., BAIER, S., HARLIZIUS, J., NIENHOFF, H., BRASE, K., ZIMMERMANN, S. & SEYBOLDT, C. 2013. Clostridium difficile genotypes in piglet populations in Germany. *J.Clin.Microbiol.,* 51**,** 3796–3803.

SHE, R. C., DURRANT, R. J. & PETTI, C. A. 2009. Evaluation of enzyme immunoassays to detect Clostridium difficile toxin from anaerobic stool culture. *Am J Clin Pathol,* 131**,** 81-4.

SHETTY, N., WREN, M. W. & COEN, P. G. 2011 The role of glutamate dehydrogenase for the detection of Clostridium difficile in faecal samples: a meta-analysis. *J Hosp Infect.,* 77**,** 1-6.

VANEK, J., HILL, K., COLLINS, J., BERRINGTON, A., PERRY, J., INNS, T., GORTON, R., MAGEE, J., SAILS, A., MULLAN, A. & GOULD, F. K. 2012. Epidemiological survey of Clostridium difficile ribotypes in the North East of England during an 18-month period. *J Hosp Infect,* 81**,** 209-12.

WALK, S. T., MICIC, D., JAIN, R., LO, E. S., TRIVEDI, I., LIU, E. W., ALMASSALHA, L. M., EWING, S. A., RING, C., GALECKI, A. T., ROGERS, M. A., WASHER, L., NEWTON, D. W., MALANI, P. N., YOUNG, V. B. & ARONOFF, D. M. 2012. Clostridium difficile ribotype does not predict severe infection. *Clin Infect Dis,* 55**,** 1661-8.

WALKER, A. S., EYRE, D. W., WYLLIE, D. H., DINGLE, K. E., GRIFFITHS, D., SHINE, B., OAKLEY, S., O'CONNOR, L., FINNEY, J., VAUGHAN, A., CROOK, D. W., WILCOX, M. H., PETO, T. E. & INFECTIONS IN OXFORDSHIRE RESEARCH, D. 2013. Relationship between bacterial strain type, host biomarkers, and mortality in Clostridium difficile infection. *Clin Infect Dis,* 56**,** 1589-600.

WASLAWSKI, S., LO, E. S., EWING, S. A., YOUNG, V. B., ARONOFF, D. M., SHARP, S. E., NOVAK-WEEKLEY, S. M., CRIST, A. E., JR., DUNNE, W. M., HOPPE-BAUER, J., JOHNSON, M., BRECHER, S. M., NEWTON, D. W. & WALK, S. T. 2013. Clostridium difficile ribotype diversity at six health care institutions in the United States. *J Clin Microbiol,* 51**,** 1938-41.

WIEGAND, P. N., NATHWANI, D., WILCOX, M. H., STEPHENS, J., SHELBAYA, A. & HAIDER, S. 2012. Clinical and economic burden of Clostridium difficile infection in Europe:a systematic review of healthcare-facility-acquired infection. . *J Hosp Infect.* , 81**,** 1–14.

XIAO, M., KONG, F., JIN, P., WANG, Q., XIAO, K., JEOFFREYS, N., JAMES, G. & GILBERT, G. L. 2012. Comparison of two capillary gel electrophoresis systems for Clostridium difficile ribotyping, using a panel of ribotype 027 isolates and whole-genome sequences as a reference standard. *J Clin Microbiol,* 50**,** 2755-60.

# Chapter 4

## Different SARS-CoV-2 haplotypes associate with geographic origin and case fatality rates of COVID-19 patients

Manisha Goyal, Katrien De Bruyne, Alex van Belkum, Brian West

**Highlights**

• Different allelic variants among 692 SARS-CoV-2 genome sequences display a statistically significant association with geographic origin and also COVID-19 case severity.

• Geographic variation is associated with both case severity and allelic variation especially in strains of Indian origin

• An apparent association between viral genotype and patient case severity is likely due to shared geographic heterogeneity, rather than a direct effect

**Summary**

The current pandemic of COVID-19 is caused by the SARS-CoV-2 virus for which many variants at the Single Nucleotide Polymorphism (SNP) level have now been identified. We show here that different allelic variants among 692 SARS-CoV-2 genome sequences display a statistically significant association with geographic origin ($p < 0.000001$) and COVID-19 case severity ($p = 0.016$). Geographic variation in itself is associated with both case severity and allelic variation especially in strains from Indian origin ($p < 0.000001$). Using an new alternative bioinformatics approach we were able to confirm that the presence of the D614G mutation correlates with increased case severity in a sample of 127 sequences from a shared geographic origin in the US ($p = 0.018$). While leaving open the question on the pathogenesis mechanism involved, this suggests that in specific geographic locales certain genotypes of the virus are more pathogenic than others. We here show that viral genome polymorphisms may have an effect on case severity when other factors are controlled for, but that this effect is swamped out by these other factors when comparing cases across different geographic regions.

**Introduction**

The SARS coronavirus-2 (SARS-CoV-2) causes COVID19 (Kadkhoda, 2020). This disease is now pandemic and it is killing hundreds of thousands of people on a global scale (e.g. (Potere et al., 2020)). Viruses, especially those with an RNA genome, have a tendency to evolve relatively rapidly during episodes of intense geographic spread. Using modern genomic sequencing technologies the genetic changes associated with global but also more local dissemination can be documented rapidly (Sekizuka et al., 2020). Also, within viral populations variants can be traced due to the quasi-species nature of the SARS-CoV-2 virus (Jary et al., 2020). For SARS-CoV-2 thousands of Single Nucleotide Polymorphisms (SNPs) have already been identified, several of which have become fixed in the more recent, geographically defined viral populations at large (Saha et al., 2020, Sapoval et al., 2020, Kaushal et al., 2020, Yang et al., 2020). Rapid regional spread of SARS-CoV-2 may lead to increased allelic variability during periods of extended transmission (Gudbjartsson et al., 2020). Although not all of these SNPs translate in amino acid variation in coding sequences (CDS), a significant portion does change the structure of important viral proteins. It is

currently not clear what the effect of such variations is on viral phenotypes (e.g. its capacity to adhere to target host cells, efficiency of invasion of host cells, rapidity of replication, disease features in infected hosts etc) also because defining such effects is usually performed in artificial in vitro models. Such models are often cumbersome, have an intrinsic infectious risk for those working with it and may not adequately represent the real-life in vivo situation (Lamers et al., 2020, Leibel et al., 2020). Modern bioinformatics tools may add flexibility to such laborious assays and are helpful in defining associations between viral genome variation and differential effects that such viral variants have during infection (Gallego et al., 2004, Ji et al., 2020).

Many physiological and clinical parameters have been described that significantly contribute to COVID-19 mortality. Among these are advanced age (Papadopoulos et al., 2021), smoking (Grundy et al., 2020) obesity (Hussain et al., 2020), diabetes (Rajpal et al., 2020), hypertension (Zaki et al., 2020), cardio-vascular problems (Mishra et al., 2020) and quite some others (Thompson et al., 2020, Williamson et al., 2020). Relatively little information is available on the contribution to disease severity and mortality by viral variability itself (Pachetti et al., 2020). A physiologically important mutation changing the amino acid sequence of the RNA-dependent RNA polymerase (RdRp) was noted but the effect on disease severity could not be assessed. Furthermore, it was shown that a 328 basepair deletion in ORF8 clinically associated with a lesser chance for developing hypoxia during COVID-19 (Young et al., 2020). Very recently however, (Toyoshima et al., 2020), (Nakamichi et al., 2020) and (Hodcroft et al., 2020) reported the first viral mutations that associated with fatality rates for COVID-19 and concluded that viral variation, together with host susceptibility and the environment co-define the course of COVID-19.

Using a novel viral typing tool, we here assess SNP-based haplotype variation in a large set of SARS-CoV-2 genome sequences, define the SARS-CoV-2 population structure and dynamics and associate these with clinical findings, including fatality rates, among patients.

## Materials and methods

### Collection of viral sequence information and database development

SARS-CoV-2 viral genome sequences were collected using the Global Initiative on Sharing Avian Influenza Data (GISAID) database which combined more than 90.000 genome sequences including phenotypic and disease-related metadata. Over 6400 of these sequences included relatively complete dossiers on patient status information. Sequences and metadata were stored, processed, and analyzed in a BIONUMERICS (v8.0) database, with a SQLite backend. Data quality assessment was performed by filtering the GISAID sequences for completeness (>29,000 bp) and by comparing genome sequences to the NC_045512 NCBI reference sequence. Genomic sequences were only analyzed when every CDS was the same length as the matching CDS in the NC_045512 reference sequence, i.e. without insertions or deletions.

**Bioinformatic analysis of viral sequences**

The BIONUMERICS SARS-CoV-2 plugin tool (bioMérieux, Applied Maths, Sint-Martens-Latem, Belgium) facilitates the processing and combined analysis of SARS-CoV-2 genomic sequences, whether downloaded from a public data repository or generated locally. The plugin tool is part of the BIONUMERICS platform and can be only used in the context of this software package. Each genomic sequence imported into BIONUMERICS was separated by the plugin tool into subsequences matching the annotated CDSs while ignoring the small fraction of intergenic regions in the NCBI reference sequence for SARS-CoV-2 (NC_045512). Next, each of these sequences was analyzed for SNPs relative to the reference sequence. SNPs were stored in the database as a character type experiment to be used for comparison and strain typing using BIONUMERICS' clustering tools (dendrograms and minimum spanning trees). SNPs were also translated, enabling SNP interpretation based on actual amino acid changes. The “haplotype”, as defined in the plugin, was determined by categorization of a set of common missense SNPs translated into amino acids (Sekizuka et al., 2020). This haplotype information was also stored in the database and displayed on the trees and networks for easy group detection.

**Tool modules**

After being downloaded from GISAID, FASTA-formatted genomic sequences were imported into the database using a dedicated sequence import routine available in BIONUMERICS. The SARS-CoV-2 plugin applied a BLAST approach to extract 26

subsequences from each genome. The subsequences of sample Wuhan-Hu-1 (NC_045512), installed automatically by the plugin, were used as reference sequences for the BLAST searches. The subsequences extracted from the genomic sequences were stored in the corresponding destination sequence type experiments. These sequence types were identified by ORF and, for ORF1, an additional Nuclear Shuttle Protein (nsp) tag. After the BLAST screening, the following detailed results were reported for each destination sequence type (Locus column): whether or not a BLAST hit was found, its position on the genome sequence (Start and Stop), sequence identity (Identity (%)) and sequence overlap (Length (%)), the length of the retrieved subsequence, the number of mismatches with the reference sequence (Mismatches) and the number of gaps (Open gaps) and length correction (if applied).

**Haplotype determination**

In the second step of the process, the haplotypes were determined for each sample. The haplotype, as defined in the SARS-CoV-2 plugin, consists of a set of high-frequency amino acid substitutions which are summarized in Table 4- 1. Three pairs of these substitutions were observed to be in linkage disequilibrium (DP/GL, YP/CL, and QT/HI). The substitutions are ordered on the basis of the date on which they first appeared, as inferred by Nextstrain (Hadfield et al., 2018) and with the most frequent ones being: S.DP.YP.QT, S.DP.CL.QT, L.DP.YP.QT, L.GL.YP.QT, L.GL.YP.HT, L.GL.YP.HI, L.GP.YP.QT and L.GP.YP.HT (see Figure 4- 1 for a review on their relative abundance among isolates of SARS-CoV-2).

**Table 4-1:** SARS-CoV-2 amino acid substitutions giving rise to haplotype variation as defined by genomic locus, position, and inferred date.

| Substitution | Locus | Codon # | Date |
|---|---|---|---|
| L - > S | ORF8 | 84 | 2020-01-12 |
| D - > G | S | 614 | 2020-01-12 |
| P - > L | ORF1b | 314 | 2020-01-13 |
| Q - > H | ORF3a | 57 | 2020-01-23 |
| T - > I | ORF1a | 265 | 2020-02-23 |
| Y - > C | ORF1b | 1464 | 2020-02-23 |
| P - > L | ORF1b | 1427 | 2020-02-23 |

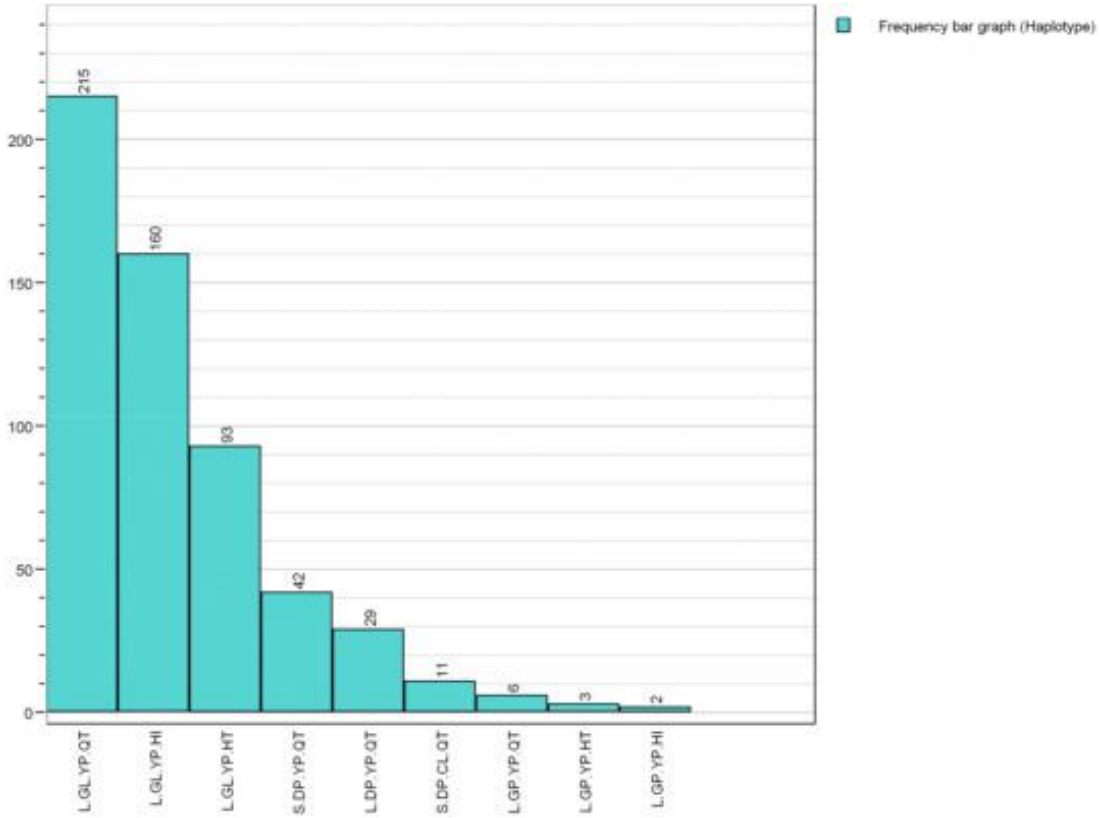



**Figure 4-1:** SARS-CoV-2 haplotype counts among samples included in this study providing adequate patient status assessment.

### SNP calculation

After extraction, the plugin screened each subsequence for SNPs by automating the built-in BIONUMERICS SNP analysis tool. The resulting SNP set was filtered based on the relaxed (non-ACGT bases allowed) SNP filtering template and the retained SNPs were stored in the SNP character experiment.

### Clustering SNP data into dendrograms

Entries to be clustered were selected based on suitability. In the first step, all selected entries were screened for the presence of the subsequences extracted in the prior processing step. Entries for which one or more subsequences are missing have an incomplete SNP character set and were excluded from the comparison. A similarity matrix was calculated based on the SNP experiment, using the categorical (differences) similarity coefficient, and displayed in the similarities panel. A dendrogram was then calculated based on the complete linkage (furthest neighbor) clustering algorithm (Sneath and Sokal, 1973). A minimum

spanning tree (MST) was then calculated in the advanced cluster analysis window of BIONUMERICS, using default priority rule settings. The SNPs stored in the SNP experiment of the selected entries were translated and the amino acids stored in the SNP_TRANSL experiment file.

### Case severity

The patient status information for each genome sequence was imported as a category (e.g. “asymptomatic”, “hospitalized”, “deceased”). Each patient's status was evaluated sometime between when the sample taken and when it was submitted, and does not necessarily reflect the case's outcome. We created a decision network in BIONUMERICS to convert each category to an integer value representing increasing case severity, on a scale from 1 to 6 (Table 4- 2).

**Table 4-2:** Patient status transformation into a numerical score of case severity.

| Patient Status | Case Severity |
|---|---|
| Asymptomatic | 1 |
| Mild case/Outpatient/Retirement home/Symptomatic | 2 |
| Alive/Released/Recovered | 3 |
| Hospitalized | 4 |
| Severe/ICU | 5 |
| Deceased | 6 |

### Statistical analysis

Tables of contingencies between two different categories (e.g. haplotypes and countries) were evaluated for unexpected frequencies with the chi-squared test. Distributions of case severity rankings across three or more categories (e.g. haplotypes or countries) were evaluated with the Kruskal-Wallace H test by ranks. Distributions of case severity rankings across two categories were evaluated with the Mann-Whitney test by sum of ranks.

### Results

We extracted 692 SARS-CoV-2 genomic sequences originating from the USA, India, Italy, France and Spain from the GISAID database. These regions were chosen for being well

represented among sequences with complete patient status metadata. The MST for these sequences shows a high degree of genotypic heterogeneity within each country although clusters representing local dissemination of closely related genotypes were obviously observed as well (Figure 4- 2). Figure 4- 2 also illustrates that strains deriving from the USA and India show global representation as well. Of note, certain types are genuinely pandemic whereas others are more geographically restricted.

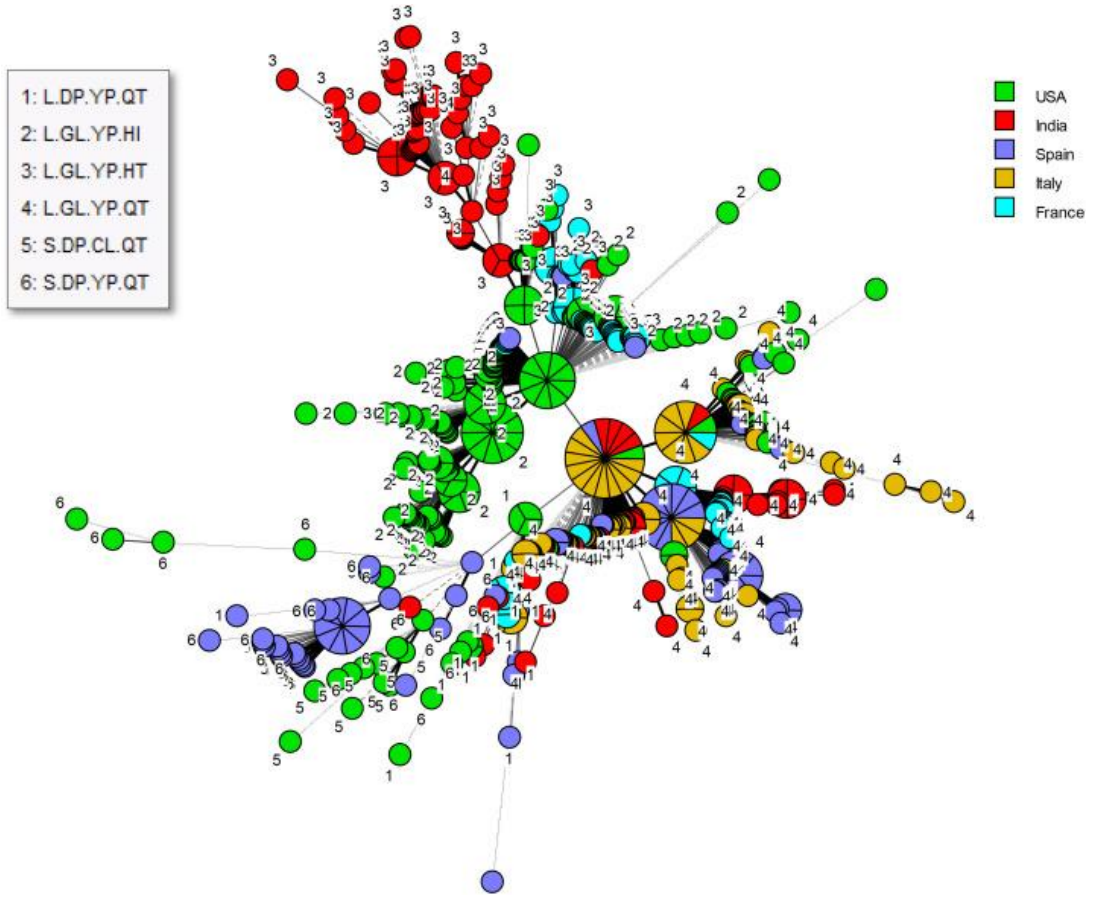



**Figure 4-2:** Minimum spanning tree for all SARS-CoV-2 genomes included in the present study. Genomes are labeled by haplotype and color-coded by country of origin.

Overall, there was a significant association between haplotype and case severity with haplotype (H = 2.360; p = 0.016743) (Figure 4- 3). There was also a strong association (H = 58.285; p = 0.000000) between case severity and country (Figure 4- 4). Furthermore, a contingency table shows a highly significant association (Chi square = 597.170, P = 0.000000) between haplotype and country (Table 4- 3). It shows that L.GL.YP.QT is widespread but predominates in Italy; that L.GL.YP.HT is found primarily in India; that S.DP.YP.QT is prominent mostly in Spain; and that L.GL.YP.HI predominates in the United States. An

examination of case severity versus haplotype within each country showed mixed results; only data from Italy and Spain showed a significant association (Table 4- 4).

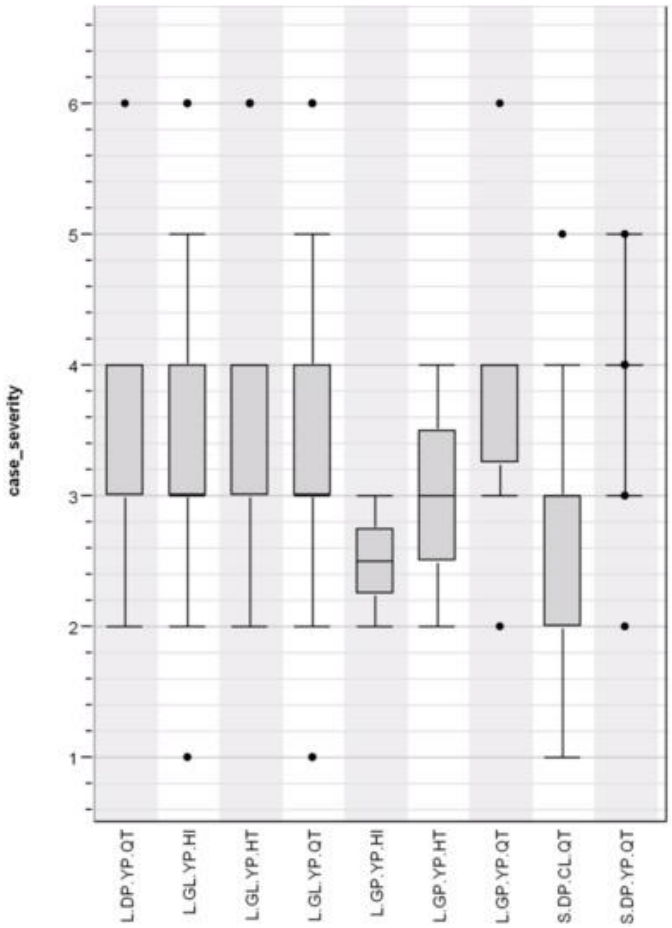



**Figure 4-3:** COVID-19 case severity by haplotype distribution (H = 2.360; p = 0.016743).

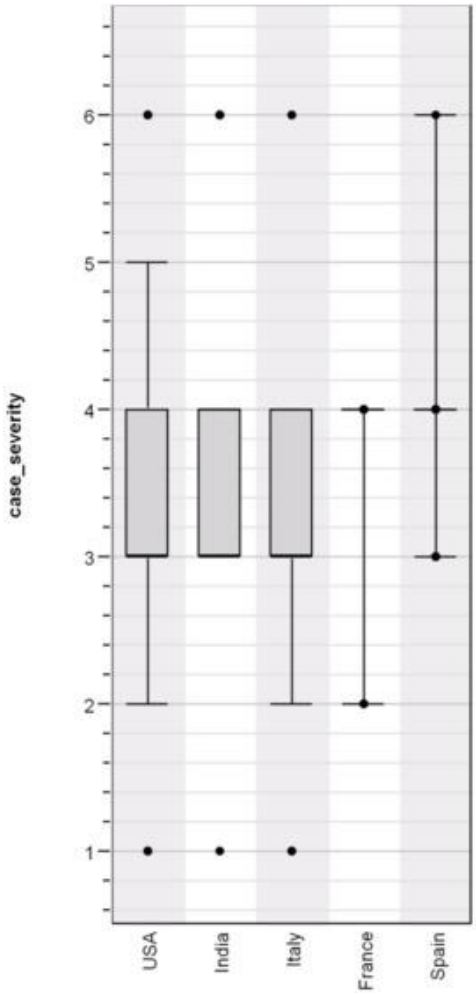



***Figure 4-4:*** *Overview of COVID-19 case severity by country of origin (H = 58.285; p = 0.000000).*

***Table 4-3:*** Contingency table for haplotype by country, with SARS-CoV-2 sequence counts shown; Chi square = 597.170, P = 0.000000.

| | **L.DP.YP.QT** | **L.GL.YP.HI** | **L.GL.YP.HT** | **L.GL.YP.QT** | **L.GP.YP.HI** | **L.GP.YP.HT** | **L.GP.YP.QT** | **S.DP.CL.QT** | **S.DP.YP.QT** |
|---|---|---|---|---|---|---|---|---|---|
| **USA** | 8 | 131 | 14 | 22 | 2 | 2 | 1 | 11 | 8 |
| **India** | 4 | 1 | 62 | 46 | 0 | 1 | 0 | 0 | 1 |
| **Spain** | 5 | 2 | 1 | 44 | 0 | 0 | 5 | 0 | 33 |
| **Italy** | 6 | 0 | 0 | 81 | 0 | 0 | 0 | 0 | 0 |
| **France** | 6 | 26 | 16 | 22 | 0 | 0 | 0 | 0 | 0 |

***Table 4-4:*** ANOVA tests on numerical case severity versus SARS-CoV-2 haplotype.

| **Country** | **Statistic** | ***P*-value** |
|---|---|---|

| Country | Statistic | *P*-value |
|---|---|---|
| USA | H = 11.222 | 0.129228 |
| India | H = 0.557 | 0.756956 |
| France | H = 2.383 | 0.496744 |
| Italy | Sum of ranks:<br>L.GL.YP.QT: 3429<br>L.DP.YP.QT: 399 | 0.023739 |
| Spain | H = 14.210 | 0.006653 |

To minimize geographic factors while maximizing genetic diversity, we selected the sequences from California for further analysis. As shown in Table 4- 5, these 133 sequences included all nine haplotypes, 20 of which were "D" types. A single CA sequence was submitted by Naval Health Research Center. A Kruskal-Wallis test by ranks did not show a statistically significant association between haplotype and case severity (Figure 4- 5). However, there was an apparent trend with regard to the D614G mutation (Figure 4- 6). By grouping the haplotypes into "D" and "G" types, a Mann-Whitney test revealed a significant association between the D614G genotypes and case severity (p = 0.031085). This is once more reflected in the MST (Figure 4- 7) where all of the deceased patients are shown to fall within the G allele group.

***Table 4-5:*** SARS-CoV-2 haplotype counts for geographic divisions.

| | L.DP.YP.QT | L.GL.YP.HI | L.GL.YP.HT | L.GL.YP.QT | L.GP.YP.HI | L.GP.YP.HT | L.GP.YP.QT | S.DP.CL.QT | S.DP.YP.QT |
|---|---|---|---|---|---|---|---|---|---|
| **California** | 8 | 76 | 14 | 18 | 2 | 2 | 1 | 6 | 6 |
| **Gujarat** | 2 | 1 | 62 | 44 | 0 | 1 | 0 | 0 | 0 |
| **Ile de France** | 6 | 26 | 16 | 22 | 0 | 0 | 0 | 0 | 0 |
| **Louisiana** | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Abruzzo** | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 0 |
| **Basque Country** | 0 | 0 | 0 | 13 | 0 | 0 | 3 | 0 | 5 |
| **Lombardy** | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 |
| **Texas** | 0 | 9 | 0 | 3 | 0 | 0 | 0 | 1 | 2 |
| **Friuli Venezia Giulia** | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 |

| | L.DP.YP.QT | L.GL.YP.HI | L.GL.YP.HT | L.GL.YP.QT | L.GP.YP.HI | L.GP.YP.HT | L.GP.YP.QT | S.DP.CL.QT | S.DP.YP.QT |
|---|---|---|---|---|---|---|---|---|---|
| **Apulia** | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 |
| **Andalusia** | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 7 |
| **Aragon** | 1 | 0 | 1 | 7 | 0 | 0 | 0 | 0 | 1 |
| **Galicia** | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 4 |
| **La Rioja** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| **Castilla** | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 5 |
| **Campania** | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| **Puerto Rico** | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 3 | 0 |
| **Lazio** | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Veneto** | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 |
| **Melilla** | 0 | 1 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| **Catalunya** | 0 | 0 | 0 | 4 | 0 | 0 | 1 | 0 | 0 |
| **Madrid** | 1 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| **Telangana** | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Navarra** | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| **Comunitat Valenciana** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| **Canarias** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **South Carolina** | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Florida** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Marche** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **None** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Montana** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

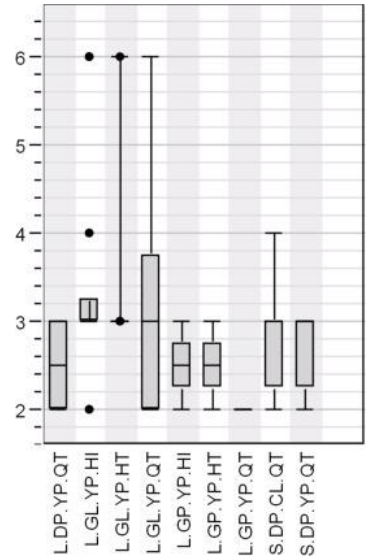



***Figure 4-5:*** *COVID case severity versus haplotype in California, USA (H = 12.514; p = 0.129694).*

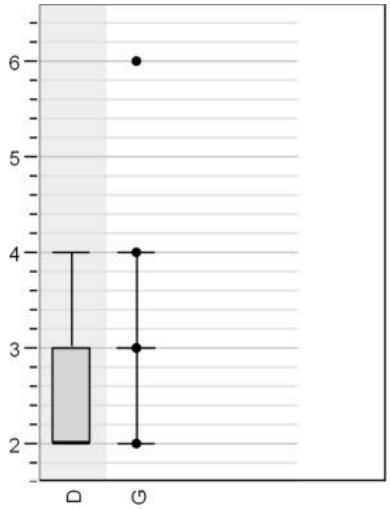



***Figure 4-6:*** *COVID case severity versus the D614G mutation (Sum of ranks: G 7913.5, D 997.5; p = 0.031085).*

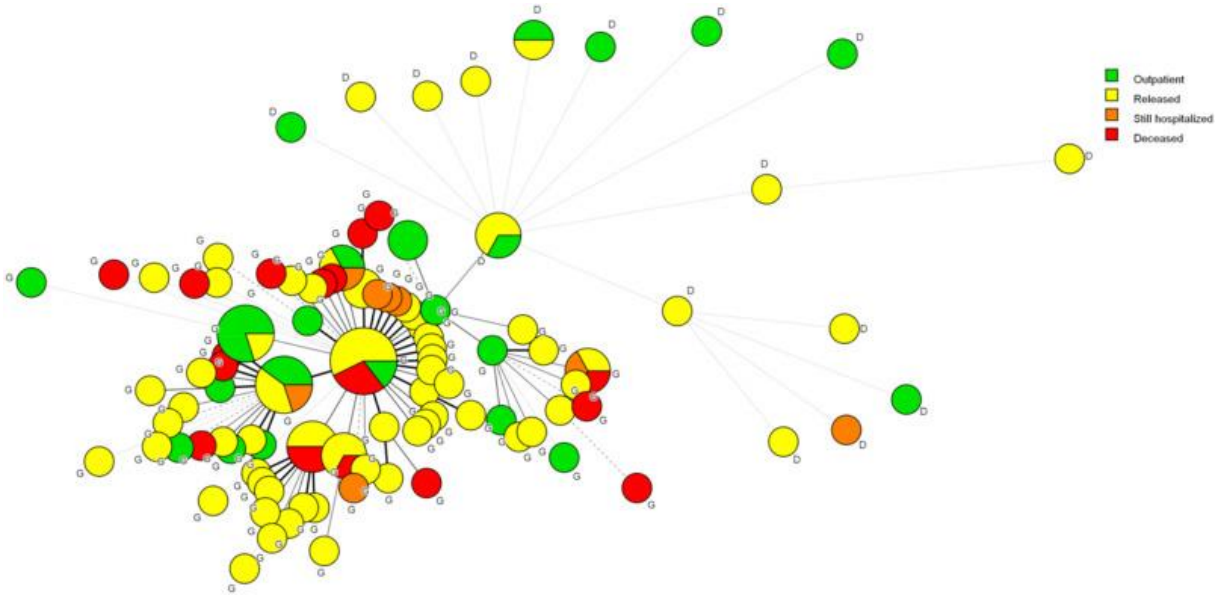



***Figure 4-7:*** *Minimum spanning tree covering haplotype diversity at the D614G level in association with disease severity. Note that deceased patients are entirely in the G cluster, as are all but one of the still hospitalized patients.*

**Discussion**

Several studies have addressed the relevance of human genetic polymorphism in severity and mortality of COVID-19 (Bosso et al., 2020, Li et al., 2020, Lu et al., 2020, McCoy et al., 2020, Asselta et al., 2020). Host variation is usually associated with pathogen adaptation and evolution. The relevance of viral variation in this respect has been studied by (Parlikar et al., 2020) who analyzed 167 SARS-CoV-2, 312 SARS-CoV, and 5 Pangolin CoV genomes to help understand their origin and evolution. The phylogeny of the subgenus Sarbecovirus confirmed the fact that SARS-CoV-2 strains evolved from their common ancestors putatively residing in bat or pangolin hosts. These authors predicted a few country-specific patterns of relatedness but failed to document any relatedness between genotypes and disease phenotypes in human patients. Two other recent publications again touch upon a lack of viral variation in the development of more or less serious disease. In the review by (Callaway et al., 2020) it is concluded that viral mutations do not contribute to mortality and that more likely than not environmental conditions have a more significant clinical impact than viral variation. (Zhang et al., 2020) conclude similarly, based on the bioinformatic analyses of experimentally defined genome sequences. In this study, the number of clinical isolates sequenced may have been a limiting factor.

We have here set out to correlate viral genotypes with host phenotypes in more detail using a large number of SARS-CoV-2 genome sequences from a broader geographic origin. We show that genotypic variants across multiple geographic regions are associated with variation in case severity. Given the likelihood that both genotype and case severity are influenced by other geographic factors, we controlled for geographic variation by focusing on one region with a relatively high degree of genotypic variation. Within this region, we showed a significant association between the D614G mutation and case severity. We also demonstrated that controlling for confounding parameters had a big effect on retrieving significant correlations between viral types and pathogenicity within patients.

The D614G mutation has received a great deal of attention with respect to its rapid global dissemination (Dearlove et al., 2020) and its significant influence on the spike protein's affinity for the ACE2 receptor. Recent studies demonstrated that in situ images of S trimer conformational changes were affected by the D614G substitution (Ke et al., 2020). This mutation abolishes a salt bridge to K854 and may reduce folding of the 833–854 loop. It has been suggested (Korber et al., 2020) that this mutation increases the virus' transmissibility, without necessarily increasing its virulence, thereby explaining its rapid spread in multiple locations. A counterargument (Grubaugh et al., 2020) has proposed that genetic drift and founder effects could also explain this pattern. More recently, the D614G mutation was identified as a marker associated with fatality rate at a countrywide level (Toyoshima et al., 2020). Our current results support these findings independently, using a completely different set of sequences and an alternative bioinformatic approach, and here show that this mutation could in fact result in increased case severity. However, we cannot rule out the possibility that transmissibility and virulence are not independent. Even if 614G is not more virulent than its D614 ancestor, ease of transmission could lead to higher viral loads in actual patients, thereby increasing the likelihood of severe cases. The polymorphisms we have identified in this project may have an effect on case severity when other factors are controlled, but that this effect is swamped out by these other factors when comparing cases across different geographic regions. Future studies should investigate the relationships among genotype, viral load, and patient outcome to sort out the underlying mechanisms.

Although this study focused on genotypes that were of particular interest at the time the data were gathered, our approach could be adapted easily to novel variants such as B.1.1.7, first observed in the UK (England, 2020). A recent update to the BIONUMERICS

SARS-CoV-2 plugin includes a tool to identify mutations relative to the reference sequence that are monomorphic for the samples of interest. For example, a set of known B.1.1.7 samples can be used to define a set of characteristic mutations, which can then be used to identify unknown samples. Once samples are characterized as variants in this way, they can be compared to other variants in terms of geography, patient outcome, and other epidemiological factors.

## Conflicts of interest

All authors are employees of bioMérieux, a company designing, developing, and selling diagnostic tests for infectious diseases. For this reason, it is impossible to provide the software used free of charge to all except for BIONUMERICS evaluation licenses and for a limited period of a month only. We do welcome collaborations in order to expand the current type of analyses and look forward to suggestions to that effect. BioMérieux marketing and sales departments had no part in the design and the written documentation of this work.

## Acknowledgements

We gratefully acknowledge Dr. Maud Tournoud (bioMérieux, Data Analytics, Grenoble, France) for editing the paper and advising on proper statistical procedures to be used.

## References

ASSELTA, R., PARABOSCHI, E. M., MANTOVANI, A. & DUGA, S. 2020. ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. *Aging (Albany NY),* 12**,** 10087.

BOSSO, M., THANARAJ, T. A., ABU-FARHA, M., ALANBAEI, M., ABUBAKER, J. & AL-MULLA, F. 2020. The two faces of ACE2: the role of ACE2 receptor and its polymorphisms in hypertension and COVID-19. *Molecular Therapy-Methods & Clinical Development,* 18**,** 321-327.

CALLAWAY, E., LEDFORD, H. & MALLAPATY, S. 2020. Six months of coronavirus: the mysteries scientists are still racing to solve. *Nature***,** 178-179.

DEARLOVE, B., LEWITUS, E., BAI, H., LI, Y., REEVES, D. B., JOYCE, M. G., SCOTT, P. T., AMARE, M. F., VASAN, S. & MICHAEL, N. L. 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proceedings of the National Academy of Sciences,* 117**,** 23652-23662.

ENGLAND, P. 2020. Investigation of novel SARS-CoV-2 variant: variant of concern 202012/01. *Public Health,* 408.

GALLEGO, O., MARTIN-CARBONERO, L., AGUERO, J., DE MENDOZA, C., CORRAL, A. & SORIANO, V. 2004. Correlation between rules-based interpretation and virtual phenotype interpretation of HIV-1 genotypes for predicting drug resistance in HIV-infected individuals. *Journal of virological methods,* 121**,** 115-118.

GRUBAUGH, N. D., HANAGE, W. P. & RASMUSSEN, A. L. 2020. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell,* 182**,** 794-795.

GRUNDY, E. J., SUDDEK, T., FILIPPIDIS, F. T., MAJEED, A. & CORONINI-CRONBERG, S. 2020. Smoking, SARS-CoV-2 and COVID-19: A review of reviews considering implications for public health policy and practice. *Tobacco induced diseases,* 18.

GUDBJARTSSON, D. F., HELGASON, A., JONSSON, H., MAGNUSSON, O. T., MELSTED, P., NORDDAHL, G. L., SAEMUNDSDOTTIR, J., SIGURDSSON, A., SULEM, P. & AGUSTSDOTTIR, A. B. 2020. Spread of SARS-CoV-2 in the Icelandic population. *New England Journal of Medicine,* 382**,** 2302-2315.

HADFIELD, J., MEGILL, C., BELL, S. M., HUDDLESTON, J., POTTER, B., CALLENDER, C., SAGULENKO, P., BEDFORD, T. & NEHER, R. A. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics,* 34**,** 4121-4123.

HODCROFT, E., ZUBER, M., NADEAU, S., COMAS, I. & GONZÁLEZ CANDELAS, F. 2020. SeqCOVID-SPAIN consortium. Stadler T, Neher RA. *Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of*.

HUSSAIN, A., MAHAWAR, K., XIA, Z., YANG, W. & SHAMSI, E.-H. 2020. RETRACTED: Obesity and mortality of COVID-19. Meta-analysis. *Obesity research & clinical practice,* 14**,** 295-300.

JARY, A., LEDUCQ, V., MALET, I., MAROT, S., KLEMENT-FRUTOS, E., TEYSSOU, E., SOULIÉ, C., ABDI, B., WIRDEN, M. & POURCHER, V. 2020. Evolution of viral quasispecies during SARS-CoV-2 infection. *Clinical Microbiology and Infection,* 26**,** 1560. e1-1560. e4.

JI, X., TAN, W., ZHANG, C., ZHAI, Y., HSUEH, Y., ZHANG, Z., ZHANG, C., LU, Y., DUAN, B. & TAN, G. 2020. TWIRLS, a knowledge-mining technology, suggests a possible mechanism for the pathological changes in the human host after coronavirus infection via ACE2. *Drug Development Research,* 81**,** 1004-1018.

KADKHODA, K. 2020. COVID-19: an immunopathological view. *MSphere,* 5**,** e00344-20.

KAUSHAL, N., GUPTA, Y., GOYAL, M., KHAIBOULLINA, S. F., BARANWAL, M. & VERMA, S. C. 2020. Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. *Pathogens,* 9**,** 565.

KE, Z., OTON, J., QU, K., CORTESE, M., ZILA, V., MCKEANE, L., NAKANE, T., ZIVANOV, J., NEUFELDT, C. J. & CERIKAN, B. 2020. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature,* 588**,** 498-502.

KORBER, B., FISCHER, W. M., GNANAKARAN, S., YOON, H., THEILER, J., ABFALTERER, W., HENGARTNER, N., GIORGI, E. E., BHATTACHARYA, T. & FOLEY, B. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell,* 182**,** 812-827. e19.

LAMERS, M. M., BEUMER, J., VAN DER VAART, J., KNOOPS, K., PUSCHHOF, J., BREUGEM, T. I., RAVELLI, R. B., PAUL VAN SCHAYCK, J., MYKYTYN, A. Z. & DUIMEL, H. Q. 2020. SARS-CoV-2 productively infects human gut enterocytes. *Science,* 369**,** 50-54.

LEIBEL, S. L., MCVICAR, R. N., WINQUIST, A. M., NILES, W. D. & SNYDER, E. Y. 2020. Generation of Complete Multi− Cell Type Lung Organoids From Human Embryonic and Patient-Specific Induced Pluripotent Stem Cells for Infectious Disease

Modeling and Therapeutics Validation. *Current protocols in stem cell biology,* 54**,** e118.

LI, Q., CAO, Z. & RAHMAN, P. 2020. Genetic variability of human angiotensin-converting enzyme 2 (hACE2) among various ethnic populations. *Molecular genetics & genomic medicine,* 8**,** e1344.

LU, R., ZHAO, X., LI, J., NIU, P., YANG, B., WU, H., WANG, W., SONG, H., HUANG, B. & ZHU, N. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The lancet,* 395**,** 565-574.

MCCOY, J., WAMBIER, C. G., VANO-GALVAN, S., SHAPIRO, J., SINCLAIR, R., RAMOS, P. M., WASHENIK, K., ANDRADE, M., HERRERA, S. & GOREN, A. 2020. Racial variations in COVID-19 deaths may be due to androgen receptor genetic variants associated with prostate cancer and androgenetic alopecia. Are anti-androgens a potential treatment for COVID-19? *Journal of cosmetic dermatology*.

MISHRA, A. K., SAHU, K. K., GEORGE, A. A. & LAL, A. 2020. A review of cardiac manifestations and predictors of outcome in patients with COVID–19. *Heart & Lung,* 49**,** 848-852.

NAKAMICHI, K., SHEN, J. Z., LEE, C. S., LEE, A., ROBERTS, E. A., SIMONSON, P. D., ROYCHOUDHURY, P., ANDRIESEN, J., RANDHAWA, A. K. & MATHIAS, P. C. 2020. Outcomes associated with SARS-CoV-2 viral clades in COVID-19. *medRxiv*.

PACHETTI, M., MARINI, B., BENEDETTI, F., GIUDICI, F., MAURO, E., STORICI, P., MASCIOVECCHIO, C., ANGELETTI, S., CICCOZZI, M. & GALLO, R. C. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of translational medicine,* 18**,** 1-9.

PAPADOPOULOS, V., LI, L. & SAMPLASKI, M. 2021. Why does COVID-19 kill more elderly men than women? Is there a role for testosterone? *Andrology,* 9**,** 65-72.

PARLIKAR, A., KALIA, K., SINHA, S., PATNAIK, S., SHARMA, N., VEMURI, S. G. & SHARMA, G. 2020. Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2. *PeerJ,* 8**,** e9576.

POTERE, N., VALERIANI, E., CANDELORO, M., TANA, M., PORRECA, E., ABBATE, A., SPOTO, S., RUTJES, A. W. & DI NISIO, M. 2020. Acute complications and mortality in hospitalized patients with coronavirus disease 2019: a systematic review and meta-analysis. *Critical care,* 24**,** 1-12.

RAJPAL, A., RAHIMI, L. & ISMAIL-BEIGI, F. 2020. Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes. *Journal of diabetes,* 12**,** 895-908.

SAHA, I., GHOSH, N., MAITY, D., SHARMA, N., SARKAR, J. P. & MITRA, K. 2020. Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. *Infection, Genetics and Evolution,* 85**,** 104457.

SAPOVAL, N., MAHMOUD, M., JOCHUM, M. D., LIU, Y., ELWORTH, R. L., WANG, Q., ALBIN, D., OGILVIE, H., LEE, M. D. & VILLAPOL, S. 2020. Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission. *BioRxiv*.

SEKIZUKA, T., ITOKAWA, K., KAGEYAMA, T., SAITO, S., TAKAYAMA, I., ASANUMA, H., NAO, N., TANAKA, R., HASHINO, M. & TAKAHASHI, T. 2020. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. *Proceedings of the National Academy of Sciences,* 117**,** 20198-20201.

SNEATH, P. H. & SOKAL, R. R. 1973. Numerical taxonomy: the principles and practice of numerical classification.

THOMPSON, J. V., MEGHANI, N., POWELL, B. M., NEWELL, I., CRAVEN, R., SKILTON, G., BAGG, L. J., YAQOOB, I., DIXON, M. J. & EVANS, E. J. 2020. Patient characteristics and predictors of mortality in 470 adults admitted to a district general hospital in England with Covid-19. *Epidemiology & Infection,* 148.

TOYOSHIMA, Y., NEMOTO, K., MATSUMOTO, S., NAKAMURA, Y. & KIYOTANI, K. 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *Journal of human genetics,* 65**,** 1075-1082.

WILLIAMSON, E. J., WALKER, A. J., BHASKARAN, K., BACON, S., BATES, C., MORTON, C. E., CURTIS, H. J., MEHRKAR, A., EVANS, D. & INGLESBY, P. 2020. Factors associated with COVID-19-related death using OpenSAFELY. *Nature,* 584**,** 430-436.

YANG, H.-C., CHEN, C.-H., WANG, J.-H., LIAO, H.-C., YANG, C.-T., CHEN, C.-W., LIN, Y.-C., KAO, C.-H., LU, M.-Y. J. & LIAO, J. C. 2020. Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. *Proceedings of the National Academy of Sciences,* 117**,** 30679-30686.

YOUNG, B. E., FONG, S.-W., CHAN, Y.-H., MAK, T.-M., ANG, L. W., ANDERSON, D. E., LEE, C. Y.-P., AMRUN, S. N., LEE, B. & GOH, Y. S. 2020. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *The lancet,* 396**,** 603-611.

ZAKI, N., ALASHWAL, H. & IBRAHIM, S. 2020. Association of hypertension, diabetes, stroke, cancer, kidney disease, and high-cholesterol with COVID-19 disease severity and fatality: A systematic review. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* 14**,** 1133-1142.

ZHANG, X., TAN, Y., LING, Y., LU, G., LIU, F., YI, Z., JIA, X., WU, M., SHI, B. & XU, S. 2020. Viral and host factors related to the clinical outcome of COVID-19. *Nature,* 583**,** 437-440.

# Chapter 5

## Whole Genome Multi-Locus Sequence Typing And Genomic Single Nucleotide Polymorphism Analysis For Epidemiological Typing Of *Pseudomonas Aeruginosa* From Indonesian Intensive Care Units

Manisha Goyal[1], Andreu Coello Pelegrin[1], Magali Jaillard[2], Yulia Saharman[3,4], Corné H.W. Klaassen[4], Henri A. Verbrugh[4], Juliëtte A. Severin[4] and Alex van Belkum[1,*]

[1]bioMérieux Open innovation and Partnerships, 3 Route du Port Michaud, 038391 La Balme Les Grottes, France

[2]bioMérieux EU Data Science, 376 Chemin de l'Orme, 69280, Marcy L'Etoile, France

[3]Department of Clinical Microbiology, Faculty of Medicine, Universitas Indonesia/ Dr. Cipto Mangunkusumo General Hospital, Jakarta, Indonesia.

[4]Department of Medical Microbiology and Infectious Diseases, Erasmus MC University Medical Center, Rotterdam, The Netherlands.

**ABSTRACT**

We have previously studied carbapenem non-susceptible *Pseudomonas aeruginosa* (CNPA) strains from intensive care units (ICUs) in a referral hospital in Jakarta, Indonesia (Pelegrin AC, 2019). We documented that CNPA transmissions and acquisitions among patients were variable over time and that these were not significantly reduced by a set of infection control measures. Four high risk international CNPA clones (sequence type (ST)235, ST823, ST357, ST446) dominated and carbapenem resistance was due to carbapenemase-encoding genes and mutations in the porin OprD. We present a more detailed genomic analysis of these four major clones.

With whole genome-based Multi Locus Sequence Typing (wgMLST) of the 4 CNPA clones, three to eleven subgroups with up to 200 allelic variants were observed for each of the CNPA clones. Furthermore, we analyzed the three largest CNPA clone clusters for the presence of Single Nucleotide Polymorphisms (wgSNP) to redefine CNPA transmission events during hospitalization. A maximum number 35350 SNPs (including non-informative SNPs) and 398 SNPs (excluding non-informative SNPs) was found in ST235, 34570 SNPs (including non-informative SNPs) and 111 SNPs (excluding non-informative SNPs) in ST357 and 26443 SNPs (including non-informative SNPs) and 61 SNPs (excluding non-informative SNPs) in ST823. SNPs that are excluding non-informative SNPs were commonly noticed in sensor-response regulator genes, however the majority of non-informative SNPs was found in conserved hypothetical proteins or in uncharacterized proteins. Of note, antibiotic resistance and virulence genes segregated according to the wgSNP analyses. A total of 11 transmission chains for ST235 strains were traceable, followed by 6 and 5 possible transmission chains for ST357 and ST823. The present study demonstrates the value of detailed whole genome sequence analysis for highly refined epidemiological analysis of *P. aeruginosa*.

## INTRODUCTION

*Pseudomonas aeruginosa* is a metabolically versatile Gram-negative bacterial species often blooming in soil and aquatic environments. It effectively colonizes the exposed surfaces of plants, animals and humans (Klockgether and Tummler, 2017, Kerr and Snelling, 2009). Being an opportunistic pathogen, *P. aeruginosa* is responsible for a broad spectrum of acute and chronic infections leading to high morbidity and mortality rates (Bedard et al., 2016, Juan et al., 2017, Jacobs et al., 2020). *P. aeruginosa* causes, amongst others, bloodstream infections in immunocompromised patients and healthcare-associated infections such as ventilator-associated pneumonia and wound infections (Lodise et al., 2007, Kerr and Snelling, 2009, Doring et al., 2011). In the USA, *P. aeruginosa* causes a total of about 51,000 deadly healthcare infections per year (Fujii et al., 2014, CDC, 2019, Health and Services, 2019). Moreover, *P. aeruginosa* is known for its potential multidrug resistance (MDR) and has become one of the most troublesome causes of a wide range of intensive care unit (ICU)-acquired infections (Moore et al., 2014). The ability to develop antibiotic resistance via both mutations and resistance gene acquisitions renders *P. aeruginosa* an increasingly problematic human pathogen (Livermore, 2002, Cabot et al., 2016, De Oliveira et al., 2020). Mutations that cause antibiotic impermeability via the loss of OprD transmembrane channels are important in antimicrobial resistance (AMR) to carbapenems (STUDEMEISTER and QUINN, 1988, Livermore, 2002, Suresh et al., 2020, Puja et al., 2020). MDR isolates require careful epidemiological tracing, both locally, nationally and globally.

Microbiological epidemiology defines patterns of distribution for pathogens such as *P. aeruginosa.* It also precisely assesses spreading of infectious diseases in a variety of populations (Gad, 2014). In practice, microbiological epidemiological analysis often begins with microbial strain characterization. Multi Locus Sequence Typing (MLST) is a commonly used classical approach for *P. aeruginosa* strain characterization, it accurately defines evolutionary descent and lineages but it lacks the necessary resolution for the precise characterization of outbreaks caused by closely related, contemporaneous bacterial isolates (Inns et al., 2015a, Ashton et al., 2016b). Several studies have evaluated the discriminatory power and concordance of different typing methods (Rumore et al., 2018b, Gateau et al., 2019b). However, high throughput whole genome sequencing (WGS) is rapidly becoming the most efficient solution for strain typing of *P. aeruginosa*, both for surveillance as for (retrospective) outbreak investigations (Kan et al., 2018b). WGS facilitates whole genome MLST (wgMLST) which displays higher discrimination than conventional MLST which is

based on the analysis of seven housekeeping genes. wgMLST reliably recognizes and quantifies the genetic links between epidemiologically related isolates within various bacterial species (Cody et al., 2013, Joensen et al., 2014b, Kovanen et al., 2014b). A recent study (Blanc et al., 2020) showed that the *P. aeruginosa* wgMLST scheme in BioNumerics™ is as discriminatory as the core genome Single Nucleotide Polymorphism (cgSNP) calling approach and is hence useful for outbreak investigations. Whole genome SNP-analysis (wgSNP) is a more advanced method of exploiting variation at the WGS level to help identify bacterial transmission dynamics and to generate useful insights into the sources and routes of infection, again for essentially all bacterial species (Bakker et al., 2011, Halachev et al., 2014a, Taylor et al., 2015).

In a prior study, relatedness of carbapenem non-susceptible *P. aeruginosa* (CNPA) strains from an Indonesian hospital was analyzed at the cgSNP level (Pelegrin et al., 2019). In the present study, epidemiological correlation between the same isolates is studied on the basis of wgMLST and wgSNP analyses. Detailed wgSNP analysis was done for the pandemic *P. aeruginosa* sequence types (ST) ST235, ST357 and ST823 to reveal exact transmission patterns among patients and between patients, and the environment.

## METHODOLOGY

### Strain collection

We have used preexisting genomic data of CNPA strains collected in two ICU's of a large referral hospital in Jakarta, Indonesia (see the dataset used by (Pelegrin et al., 2019, Saharman et al., 2019 ). For each patient involved the dates of admission and discharge from the ICU were available, as well as the date of all cultures taken during ICU stay. All patients were screened for CNPA on admission, at discharge and weekly if their stay exceeded 7 days. Patients were additionally sampled upon clinical indication. Patients were enrolled in two separate episodes, before and after an infection prevention and control intervention. In the pre-intervention period ICU personnel was screened once and the ICU environment was screened twice

For the present study all CNPA genome sequences were assembled and analyzed using BioNumerics™ (Applied Maths, bioMérieux, Belgium). Antibiotic susceptibility testing

(AST) of CNPA strains was performed as described by (Pelegrin et al., 2019) using VITEK2 (bioMérieux).

**MLST and wgMLST analysis**

Classical MLST typing is based on polymorphisms in seven housekeeping genes stored in the *P. aeruginosa* pubMLST database (http://pubmlst.org). Although MLST analysis for CNPA was already published previously (Pelegrin et al., 2019), here we have repeated the analysis with an updated version of the pubMLST database in order to compare up-to-date MLST with the current wgMLST analysis. wgMLST typing of 237 CNPA genomes was performed using BioNumerics™. For wgMLST typing, fully functional and well curated schemes have been developed and maintained for many important pathogens including *P. aeruginosa* by BioNumerics™ plugins (www.applied-maths.com/applications/wgmlst). A total of 15,143 genes and other genetic elements were used to assign wgMLST types to the isolates in the CNPA collection. Allelic differences between isolates sharing the same MLST group were calculated and sub-groupings within the MLST groups were visualized as UPGMA based phylogenetic trees.

**wgSNP analysis**

Using the BioNumerics™ wgSNP application (www.applied-maths.com/applications/whole-genome-snp-analysis) wgSNPs were identified and mapped on CNPA genomes using the *P. aeruginosa* reference genome PAO-1 (Stover et al., 2000) and NCBI Reference Sequence: NC_002516.2 (Subedi et al., 2019, Pelegrin et al., 2019). Functional annotation was performed for each SNP. The option of SNP filtering was chosen during the analysis in order to remove ambiguous bases, unreliable bases and gaps. Because of these filters number of SNPs were dropped down from many thousands to upto 3-4 thousands which is still due to the inclusion of non-informative SNPs. Since the non-informative SNPs are present in all the isolates, they don't tell us anything about the genetic relationship among the isolates therefore non-informative SNP filtering (also called strict filtering) was used to get only informative SNPs which were left under 100 in numbers. Phylogenetic trees were built on the basis of wgSNPs and correlation studies were performed to identify the links among wgSNPs, wgMLST, patient characteristics, sample type, and the

resistomes and virulomes of each CNPA isolate. Resistomes and virulomes were defined using the command line script of Torsten Seemann called Abricate and which is available at Github (https://github.com/tseemann/abricate) (Zankari et al., 2012, Chen et al., 2016). Moreover, detailed maps of functional point mutations in the three most prominent classical MLST groups (ST823, ST235 and ST357) were made and possible transmission chains and routes were traced. To do so, a SNP threshold was calculated on the basis of the similarity matrix for all the CNPA strains, calculated via the SNP analysis in BioNumerics$^{TM}$. The epidemiological link provided by the clinical data collected from the hospital between patients we were able to set the SNP threshold which allowed us to sort the observed distances in the SNP distance matrix into two categories: related and not related. Here sensitivity and specificity are diagnostically equally important and desirable. Therefore, the Youden's index in conjunction with receiver operating characteristic (ROC) curve was used to indicate the performances of the different SNP cutoff values. The optimal SNP cutoff value was computed using different Youden indicators where sensitivity, accuracy, specificity and Youden's Index were found to be at their maximum.

## RESULTS AND DISCUSSION

### MLST vs wgMLST

MLST was designed primarily for the purpose of defining global bacterial phylogeny by sequencing internal fragments of seven housekeeping genes (Urwin and Maiden, 2003, Jolley et al., 2018, Maiden et al., 1998). Our set of 237 CNPA strains included 4 dominant MLST groups: ST235 (74 isolates) followed by ST357 (72 isolates), ST823 (47 isolates) and ST446 (18 isolates). The remaining 26 isolates belonged to 16 different STs (Figure 5- 1, shown for reasons of comparison with the genomic methods). Our current MLST results were in complete agreement with those presented by Pelegrin et al. 2019 Still, WGS has become the preferred method for studying the molecular epidemiology of bacterial species, clearly providing discriminatory power exceeding that of classical MLST (Kovanen et al., 2014b, Pearce et al., 2018). wgMLST analysis allows genome comparisons and recognition of evolutionary subgroups of genetically related isolates within the same classical STs, allowing more refined tracing of the origin of outbreaks and individual infections (Cody et al., 2013, Moura et al., 2016). In the present study we identified subgroups within the four major STs (ST235, ST357, ST823, ST446), but also within minor STs (Figure 5- 2). Within ST235 the

number of allelic variants between isolates ranged between 0 to 200 (11 subgroups) whereas for ST357 it was 0 to 59 (8 subgroups), for ST823 it was 0 to 39 (8 subgroups) and for ST446 it was 0 to 20 (3 subgroups). Interestingly, ST235 seemed to contain three different lineages of strains that separated earlier compared to the subgroups detected in the other STs of CNPA. Subgroupings were not specific to patients or clinical sample type, but appeared to be independent and at random (Figure 5- 2). More genetic diversity in ST235 can be seen in wgMLST tree as compared to the other STs however this might be due to a higher mutation rate in this clade as compared to that of other lineages or indeed, by chance due to an earlier occurrence of diversity within this clade. The main finding here is that wgMLST shows significantly enhanced resolving power as compared to classical MLST. Still there is excellent concordance between the two methods since there was never any mixing of classical MLST groups at the level of wgMLST groups (Blanc et al., 2020). A study conducted by (Stanton et al., 2020) demonstrated how a core genome MLST (cgMLST) scheme provided enhanced resolution over traditional MLST, pulsed-field gel electrophoresis (PFGE), and single-nucleotide variant (SNV) assessment to analyze individual outbreaks. That study included core genes those were common to all strains of *P. aeruginosa*. In contrast, wgMLST also covers highly variable elements such as repetitive genes and pseudogenes, depending upon the microbial species studied (Moura et al., 2016). However, clustering of strains based on either the cgMLST or wgMLST can provide a detailed perspective of the taxonomy, epidemiology and evolution of bacterial populations (McNally et al., 2016).

**wgSNP distribution in CNPA isolates**

wgSNP analysis represents an effective method for characterizing pathogenic bacterial strains and for detecting outbreak events (Bakker et al., 2011, Taylor et al., 2015, Schurch et al., 2018). Recent studies successfully demonstrated the capability of wgSNP-based genotyping to reveal recombination events in *Streptococcus pneumoniae*, *Staphylococcus aureus* and *Cronobacter sakazakii* (Roe et al., 2016, Cowley et al., 2018, Yin and Yau, 2018, Yong et al., 2018). To ensure the accuracy and consistency of SNP-defined outbreak analysis, essential parametric measures such as minimum coverage and distances allowed between SNPs and exclusion of non-informative SNPs must be applied (Bakker et al., 2011). In the present study, false SNPs generated due to sequencing or assembling errors were filtered. Retained SNPs including non-informative SNPs were found scattered all across the CNPA genomes. However,

after strict filtering of non-informative SNPs, SNP counts fall down into the hundreds only (Supplementary Table 1 to 3). Evolutionary relationships between CNPA strains based on wgSNP analyses can be shown in a phylogenetic tree along with their more descriptive epidemiological data and their resistomes and virulomes (Figure 5- 3). It is noteworthy that in the phylogenetic tree, clade ST235 is apparently more homogenous than other STs in terms of their resistome as well as virulome (Figure 5- 3). Detailed wgSNP analysis and annotation was performed within only the three most dominant CNPA clones (ST235, ST357 and ST823) present in the collection. Previous studies reported ST357 and ST235 clones to be prevalent across the globe and to present a high risk for invasive infection (Treepong et al., 2018, Mihara et al., 2020). The recent emergence of ST823 and ST446 causing outbreaks in many countries highlighted the importance of evaluating epidemiological trends for these clones (Zowawi et al., 2018, Pelegrin et al., 2019, Tada et al., 2019). In the present study a total of number of SNPs (including non-Informative SNPs) were ranging from 35350 in ST235, 34570 in ST357 and 26443 in ST823. Filtered and informative SNPs ranged from 398 SNPs within ST235 followed by 111 in ST357 to 61 in ST823. All point mutations, their positions and their respective functional annotations are summarized in Supplementary Tables 4 to 6. Interesting fact was that out of 12 possible SNP types (based on the availability of the 4 [A, T G and C] bases) only two, C>T and G>A, SNPs were dominant in all the three clones of CNPA (Figures 5- 4A to 4C). These two SNP types should be further investigated in order to clarify the significance of their predilection in, for instance, genetic adaptation to changes in the environment where *P. aeruginosa* is residing. SNPs were regularly found in transcription regulators and sensor-response regulator hybrids in all the three clones of CNPA strains. Non-informative SNPs were those which were present throughout the particular MLST group and very high in numbers however the occurrence of informative SNPs was different, random and limited to selective number of CNPA strains in the dataset (Supplementary Tables 4 to 6). Common point mutations (excluding non-informative SNPs) that were shared by both ST357 and ST235 include those in cytochrome C550 (A>G), the oprD porin (C>T), ABC transporters (T>A), MFS transporters (C>T) and a two-component sensors (G>A). The results presented here underscore that wgSNP typing has a higher resolution than wgMLST and that functional information on individual gene variation can be derived from the data. Of note, a large diversity of antimicrobial resistance genes and virulence genes segregated according to wgSNP analyses (Figure 5- 3).

**Transmission dynamics of CNPA clones**

By analyzing SNPs in CNPA isolates within the three major ST groups, we have determined the likelihood of their transmission from one patient to another in the setting of the two ICUs in the Indonesian hospital where the clinical part of our study took place. In this study, an optimal SNP cutoff value of <4 SNPs was calculated for the CNPA dataset using Youden indicators (Figure 5- 5). The Youden's indicator was at a maximum for a threshold of 4 SNP (sensitivity 0.86, specificity 0.96, accuracy 0.96, Youden index 0.82). As reported, individual strains of *P. aeruginosa* can be considered genetically indistinguishable if their genome sequences differ less than 3-5 SNPs (Quick et al., 2014, Parcell et al., 2018, Pelegrin et al., 2019). Previously, (Pelegrin et al., 2019) reported 50 strain acquisition events in this cohort of ICU patients on the basis of genomic proximity (at threshold < 5 SNPs) and clues from clinical data. Using the optimal SNP cutoff of 4 SNPs, we have now re-traced those acquisition events in more detail and further elucidated the chains of transmission. Genetically indistinguishable strains isolated from patients without overlapping hospitalization periods were considered as possibly originating from the same source only if the time difference between the hospitalizations, i.e. between the departure of one patient and the admission of another, was not more than 16 months (Kramer et al., 2006). These events were defined as healthcare-associated transmissions and were traced on the basis of strictly filtered SNPs (excluding non-informative SNPs) only. In the group of 36 patients harboring a ST235 strain, 5 were already carrying "their" strain at the time of admission to the ICU. The remaining patients acquired a ST235 strain during their hospitalization period. On the basis of genomic identity and the time of admission in the ICU, we found 11 possible chains of transmission including five within the ER-ICU and three in the adult-ICU; the remaining 3 events were probably inter-ICU transmissions (Figure 5- 6). In ST357, 34 patients (of which 11 were positive on admission) were involved in 7 transmission chains (Figure 5- 7), and five potential chains of transmission were detected among 25 patients harboring ST823 clones (of whom eight carried the strain at the time of their admission). In this latter group only one transmission chain occurred within the ER-ICU, two within the adult-ICU and two involving both ICUs (Figure 5- 8). Thus, where (Pelegrin et al., 2019) only presented qualitative data regarding possible transmissions and the overall number of acquisition events, we here reveal possible chains of transmission of CNPA strains between patients. Importantly, wgSNP analysis also allows for a better characterization of strains already carried by patients at the time of their hospitalization.

## CONCLUSION

We here show that wgMLST and wgSNP analyses provide enhanced resolution for the epidemiological typing of strains of *P. aeruginosa*. The use of WGS data will provide typing schemes of high discrimination capacity and, depending on the density of sampling, allow for more precise mapping of the flow of *P. aeruginosa* going through susceptible patient cohorts. This should in the end help improve infection prevention.

## ACKNOWLEDGEMENTS

This research was supported and funded by bioMérieux, France, and the European Union's Horizon 2020 research and innovation program entitled Viral and Bacterial Adhesin Network Training (ViBrANT) under Marie Skłodowska-Curie Grant Agreement No. 765042.

## CONFLICT OF INTEREST

During this study MG, MD and AvB were employees of bioMérieux, a company designing, developing, and marketing tests in the domain of infectious diseases. The company was not involved in the design of the current study and the opinions expressed are those of the authors and may be different from formal company opinions and policies.

## Figures

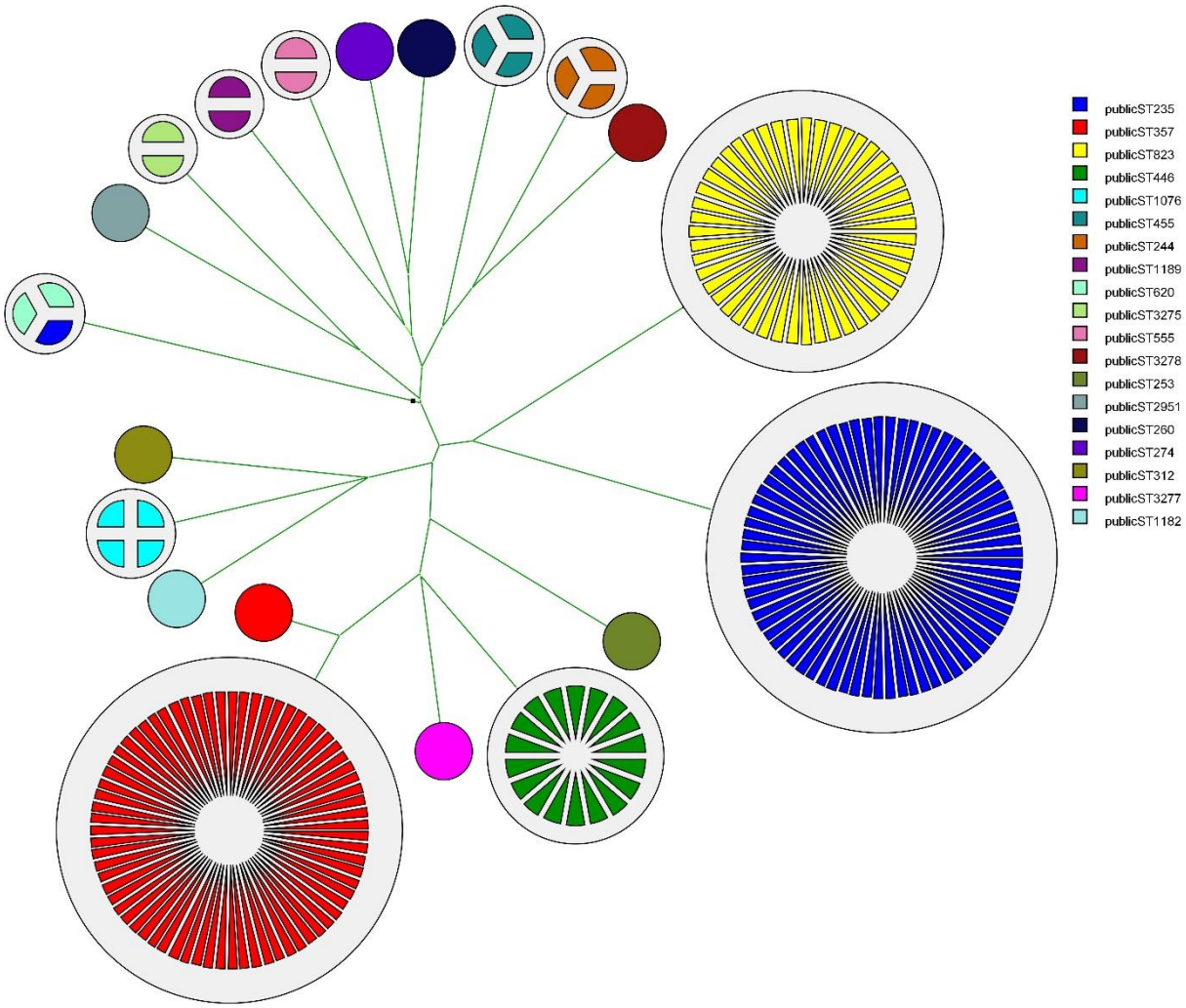



***Figure 5-1:*** *Classical MLST-based phylogenetic tree showing the evolutionary relationship between different CNPA sequence types (ST), each indicated by a different a color and provided with its ST number. Number of partitions in each cluster showing the number of strains in that group. Note that ST446, ST357, ST823 and ST235 represent the largest clonal clusters. A similar illustration was presented by Pelegrin et al (2019).*

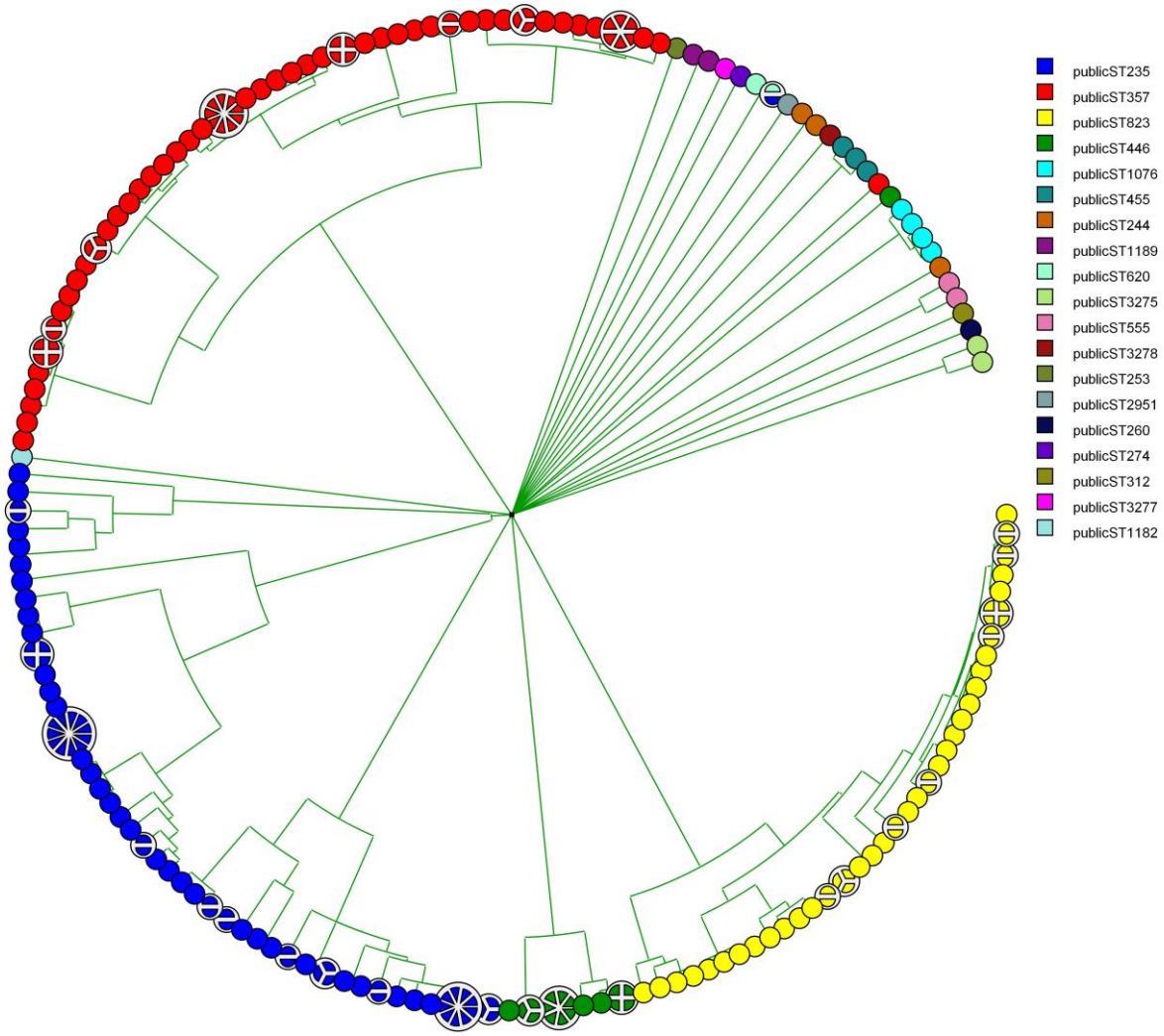



***Figure 5-2:*** *Phylogenetic tree showing CNPA relatedness based on wgMLST. Subgrouping with in each ST (denoted by different colors) is labeled by the original source (patient ID or environmental source) from which these strains have been isolated.*

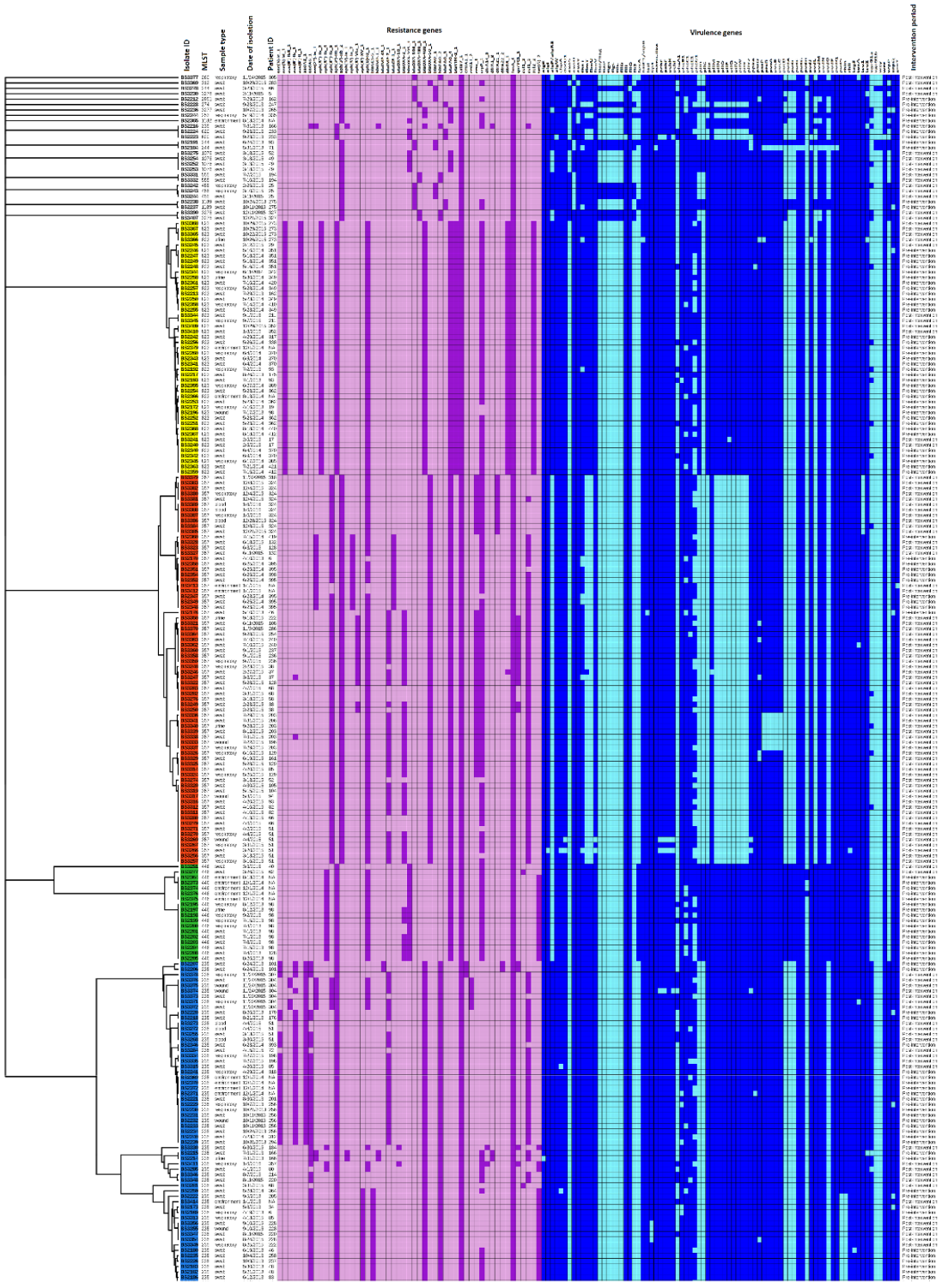



***Figure 5-3:*** *wgSNP based phylogenetic relationship between CNPA isolates. Major MLST groups are shown with different colors (yellow: ST823; red: ST235; green: ST446 and blue: ST357). Resistance and virulence genes are presented in the form of heat maps with purple and blue color ranges. Epidemiological and clinical data includes isolate ID, MLST, date and source of isolation (patients, environment and sample type) and intervention period are also mentioned along with the tree.*

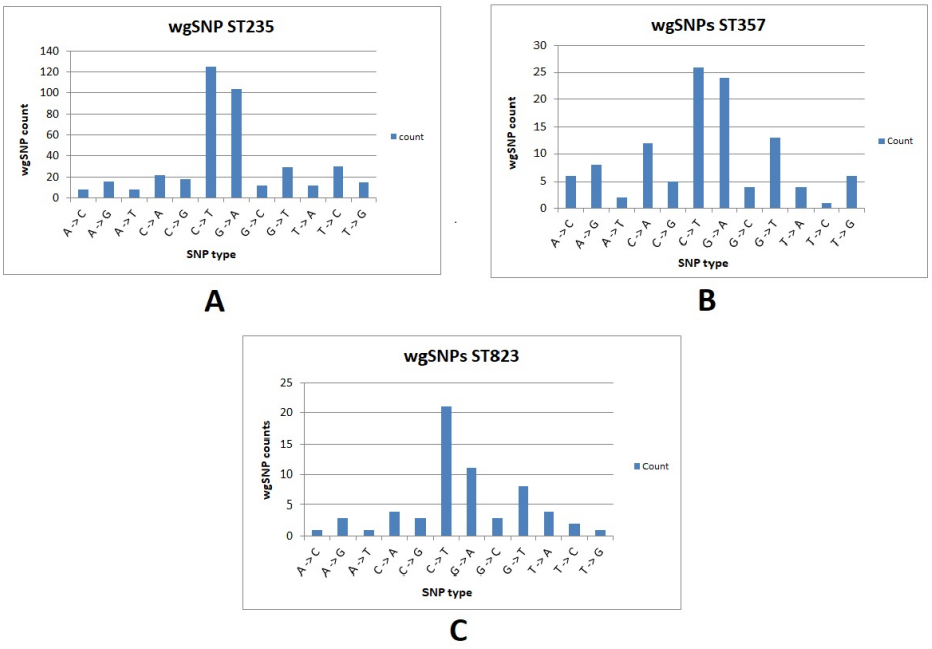



***Figure 5-4:*** *Number of different SNP types with reference to PAO1 strain of* P. aeruginosa *is illustrated for CNPA clone ST235 (A), clone ST357 (B) and clone ST823 (C).*

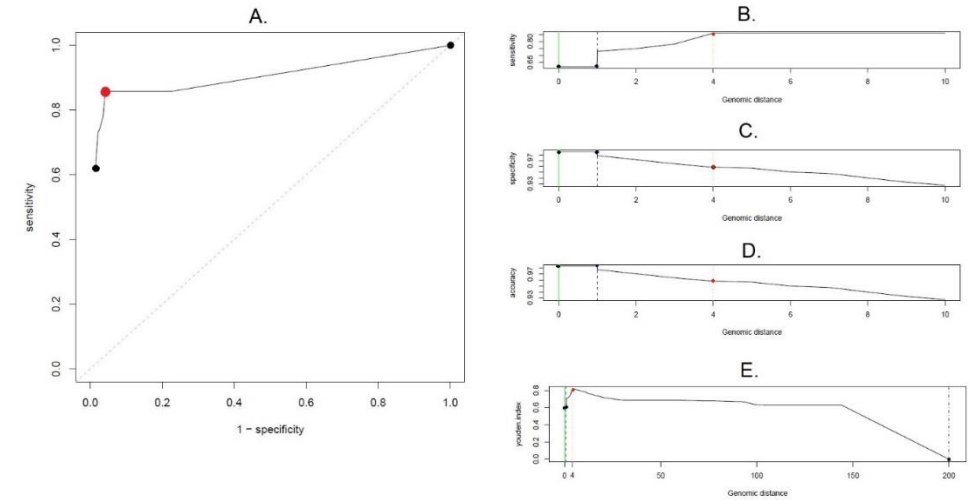



***Figure 5-5:*** *Different Youden indicators calculated using similarity matrix of CNPA strains, generated during SNP analysis. In Figure A. an ROC curve showing the relationship between clinical sensitivity and specificity for every possible SNP cut-off. Here an optimal point is represented with red colored dot. SNP cutoff values (Genomic distances) are shown on horizontal axis and different statistical parameters or indicators like Sensitivity, Specificity, Youden's Index and accuracy are shown on vertical axis in Figure B to E respectively. Based on all the above mentioned indicators genomic distance of 4 SNPs was chosen as overall optimal SNP cutoff value and is highlighted with red colored dot on each graph.*

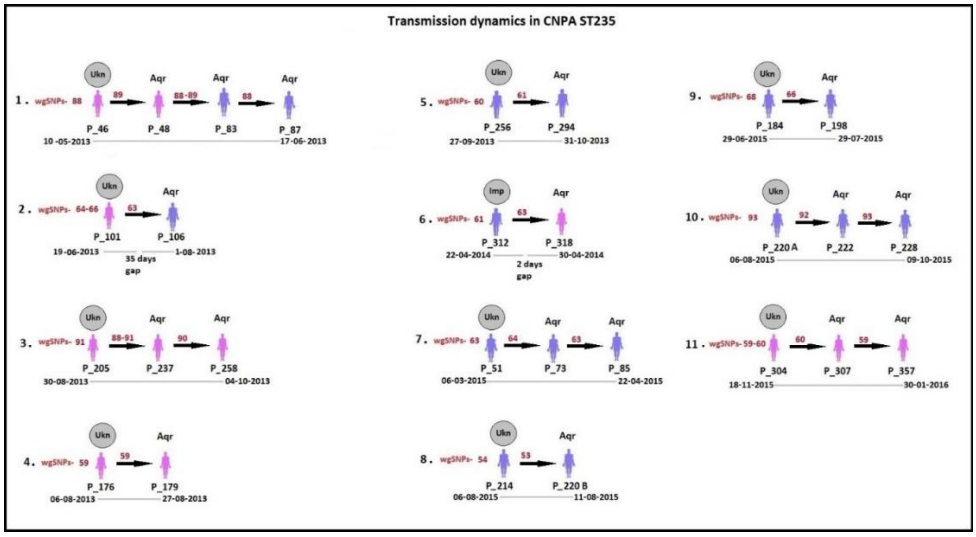



***Figure 5-6:*** *Potential transmissions of CNPA ST235 isolates among the patients (their ID given as 'P_') are shown in the figure. Pink colored patients are from adult-ICU and purple colored patients are from ER-ICU. Number of wgSNPs are shown in red colored text for each transmission event. All these transmissions are arranged in ascending order according to their time of admission to the ICU and their sample collection dates from the year 2013 to 2015. The grey circle above the first patient in each transmission event denotes that the CNPA strain was either imported (Imp) from outside at the time of admission or acquired from an unknown (Ukn) source within the ICU. Other patients in each transmission chain acquired (Aqr) these clones. The hospitalization time line of the patients is shown below each transmission chain. The gap in the time line depicts that there is a time difference between discharge of one patient and admission of another to the ICU.*

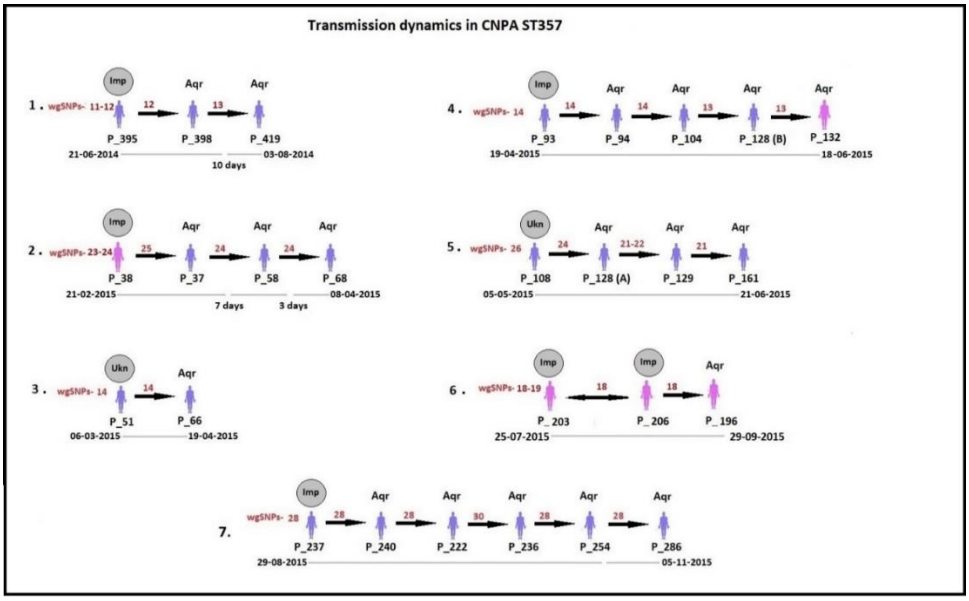



***Figure 5-7:*** *Potential transmissions of CNPA ST357 isolates among the patients are shown. Pink colored patients (their ID given below as 'P_') are from adult-ICU and purple colored patients are from ER-ICU. Numbers of wgSNPs are shown in red for each transmission event. All these transmissions are arranged in ascending order according to their time of admission*

*to the ICU and their sample collection date from the year 2013 to 2015. The grey circle above first patient in each transmission event denotes that the CNPA strain was either imported (Imp) from outside at the time admission or acquired from an unknown (Ukn) source within the ICU. Other patients in each transmission chain acquired (Aqr) these clones. Hospitalization time line of the patients is shown below each transmission chain. The gap in the time line depicts that there is a time difference between discharge of one patient and admission of another to the ICU. 6$^{th}$ Transmission chain shows an unusual situation where two patients (203 and 206) possibly the part of a transmission event but both of them were already imported with ST235 strain. Therefore the probability of transmission could be to or from P_203 or P_206 during their overlapping days of stay in the ICU.*

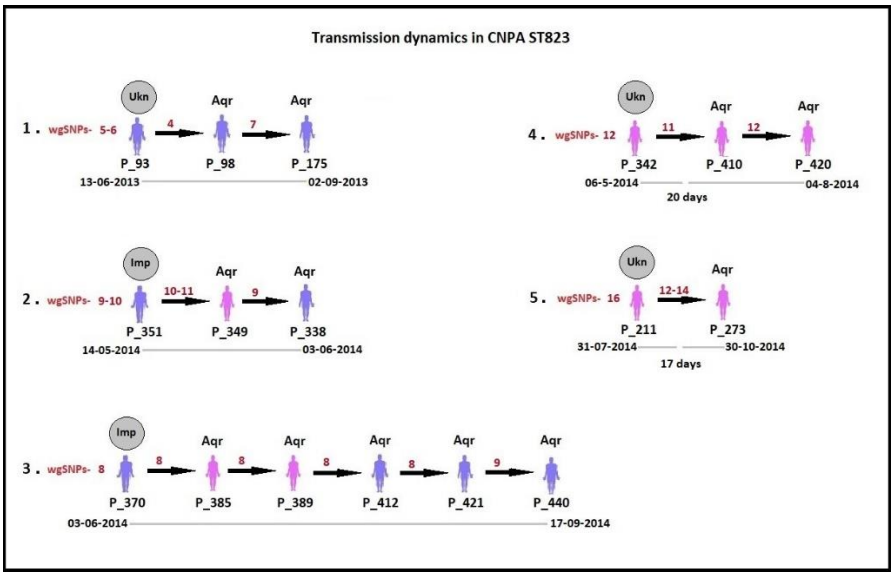



***Figure 5-8:*** *Potential transmissions of CNPA ST823 isolates among the patients (their ID given below as 'P_'). Pink colored patients are from adult-ICU and purple colored patients are from ER-ICU. Number of wgSNPs are shown in red colored text in each transmission event. All these transmissions are arranged in ascending order according to their time of admission to the ICU and their sample collection date from the year 2013 to 2015. The grey circle above first patient in each transmission event denotes that CNPA strain either imported (Imp) from outside at the time admission or acquired from unknown (Ukn) source within the ICU. Other patients in each transmission chain acquired (Aqr) these clones. Hospitalization time line of the patients is shown below each transmission chain. The gap in the time line depicts that there is a time difference between discharge of one patient and admission of another to the ICU.*

**Supplementary Table 1:** Number of wgSNPs identified in ST235 along with metadata information of strains.

| Strain ID | Isolation ID | Total SNP count | Retained SNPs (including Non informative SNPs) | Retained SNPs with strict filters (without non-informative) | Acquired/Imported | MLST PubMLST ST | Sample type | Phase | Sample collection date | Virulence genes numbr |
|---|---|---|---|---|---|---|---|---|---|---|
| BS2169_R1.fastq | 6B | 439406 | 35039 | 87 | acquired | Cluster 67 | sputum | Pre-intervention | 4/19/2013 | 216 |
| BS2173_R1.fastq | 34 | 439440 | 35038 | 86 | acquired | Cluster 67 | rectal | Pre-intervention | 5/8/2013 | 215 |
| BS2179_R1.fastq | 46C | 466917 | 35040 | 88 | acquired | Cluster 22 | throat | Pre-intervention | 5/20/2013 | 217 |
| BS2177_R1.fastq | 46E | 462298 | 35040 | 88 | acquired | Cluster 22 | rectal | Pre-intervention | 6/10/2013 | 217 |
| BS2175_R1.fastq | 46B | 466689 | 35040 | 88 | acquired | Cluster 22 | sputum | Pre-intervention | 5/20/2013 | 216 |
| BS2174_R1.fastq | 46D | 466037 | 35040 | 88 | acquired | Cluster 22 | sputum | Pre-intervention | 6/3/2013 | 216 |
| BS2178_R1.fastq | 46F | 466644 | 35041 | 89 | acquired | Cluster 22 | rectal | Pre-intervention | 6/19/2013 | 217 |
| BS2180_R1.fastq | 46G | 466889 | 35044 | 92 | acquired | Cluster 22 | throat | Pre-intervention | 6/19/2013 | 215 |
| BS2181_R1.fastq | 48C | 465665 | 35040 | 88 | acquired | Cluster 22 | throat | Pre-intervention | 5/30/2013 | 217 |
| BS2182_R1.fastq | 48A | 466050 | 35041 | 89 | acquired | Cluster 22 | throat | Pre-intervention | 5/21/2013 | 217 |
| BS2183_R1.fastq | 48B | 466874 | 35042 | 90 | acquired | Cluster 22 | rectal | Pre-intervention | 5/30/2013 | 217 |
| BS3255_R1.fastq | 51A | 474842 | 35015 | 63 | acquired | Cluster 18 | rectal | Post-intervention | 3/13/2015 | 220 |
| BS3268_R1.fastq | 51G | 470748 | 35015 | 63 | acquired | Cluster 18 | blood | Post-intervention | 3/30/2015 | 220 |
| BS3272_R1.fastq | 51K | 472425 | 35015 | 63 | acquired | Cluster 18 | blood | Post-intervention | 4/4/2015 | 218 |
| BS3273_R1.fastq | 51L | 471755 | 35015 | 63 | acquired | Cluster 18 | blood | Post-intervention | 4/4/2015 | 219 |
| BS3281_R1.fastq | 68A | 449897 | 35037 | 85 | acquired | Cluster 74 | rectal | Post-intervention | 3/31/2015 | 219 |
| BS3284_R1.fastq | 72 | 470096 | 35016 | 64 | acquired | Cluster 58 | rectal | Post-intervention | 4/15/2015 | 219 |
| BS3285_R1.fastq | 80 | 436652 | 35006 | 54 | acquired | Cluster 62 | rectal | Post-intervention | 4/1/2015 | 218 |
| BS2185_R1.fastq | 83A | 466489 | 35040 | 88 | acquired | Cluster 22 | sputum | Pre-intervention | 6/11/2013 | 216 |
| BS2186_R1.fastq | 83B | 466953 | 35041 | 89 | acquired | Cluster 22 | throat | Pre-intervention | 6/12/2013 | 217 |
| BS3315_R1.fastq | 85C | 465549 | 35015 | 63 | imported | Cluster 53 | throat | Post-intervention | 4/20/2015 | 218 |
| BS3313_R1.fastq | 85A | 438134 | 35041 | 89 | imported | Cluster 71 | sputum | Post-intervention | 4/13/2015 | 217 |
| BS2187_R1.fastq | 87 | 466459 | 35040 | 88 | acquired | Cluster 22 | rectal | Pre-intervention | 6/13/2013 | 216 |
| BS2194_R1.fastq | 95 | 463870 | 35040 | 88 | imported | Cluster 22 | throat | Pre-intervention | 6/17/2013 | 217 |
| BS2206_R1.fastq | 101A | 472478 | 35016 | 64 | acquired | Cluster 19 | rectal | Pre-intervention | 6/24/2013 | 219 |
| BS2207_R1.fastq | 101B | 471266 | 35018 | 66 | acquired | Cluster 19 | throat | Pre-intervention | 6/24/2013 | 219 |
| BS2211_R1.fastq | 141C | 464806 | 35040 | 88 | imported | Cluster 22 | rectal | Pre-intervention | 7/23/2013 | 217 |
| BS2210_R1.fastq | 141A | 464999 | 35040 | 88 | imported | Cluster 69 | throat | Pre-intervention | 7/16/2013 | 217 |
| BS2209_R1.fastq | 141B | 465472 | 35040 | 88 | imported | Cluster 22 | drain | Pre-intervention | 7/16/2013 | 217 |
| BS2215_R1.fastq | 166B | 466125 | 35015 | 63 | imported | Cluster 21 | throat | Pre-intervention | 7/31/2013 | 217 |
| BS2214_R1.fastq | 166C | 468405 | 35015 | 63 | imported | Cluster 21 | urine | Pre-intervention | 7/31/2013 | 218 |
| BS2218_R1.fastq | 176A | 470567 | 35011 | 59 | acquired | Cluster 57 | rectal | Pre-intervention | 8/21/2013 | 219 |
| BS2220_R1.fastq | 179 | 470915 | 35011 | 59 | acquired | Cluster 57 | rectal | Pre-intervention | 8/26/2013 | 218 |
| BS3330_R1.fastq | 184 | 430478 | 35020 | 68 | imported | Cluster 64 | rectal | Post-intervention | 6/30/2015 | 219 |
| BS3334_R1.fastq | 198A | 466080 | 35018 | 66 | imported | Cluster 17 | sputum | Post-intervention | 7/27/2015 | 219 |
| BS3335_R1.fastq | 198B | 469827 | 35018 | 66 | imported | Cluster 17 | rectal | Post-intervention | 7/27/2015 | 217 |
| BS2221_R1.fastq | 201 | 467564 | 35010 | 58 | imported | Cluster 54 | throat | Pre-intervention | 8/30/2013 | 219 |
| BS2222_R1.fastq | 205 | 467255 | 35043 | 91 | acquired | Cluster 65 | rectal | Pre-intervention | 9/5/2013 | 216 |
| BS3346_R1.fastq | 214 | 469014 | 35006 | 54 | imported | Cluster 61 | rectal | Post-intervention | 8/7/2015 | 220 |
| BS3348_R1.fastq | 220B | 466513 | 35005 | 53 | acquired | Cluster 61 | throat | Post-intervention | 8/11/2015 | 219 |
| BS3347_R1.fastq | 220A | 435261 | 35045 | 93 | acquired | Cluster 72 | rectal | Post-intervention | 8/11/2015 | 217 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| BS3349_R1.fastq | 222A | 438330 | 35044 | 92 | acquired | Cluster 68 | sputum | Post-intervention | 8/25/2015 | 217 |
| BS3354_R1.fastq | 228A | 441314 | 35045 | 93 | acquired | Cluster 23 | throat | Post-intervention | 8/24/2015 | 215 |
| BS3355_R1.fastq | 228B | 438345 | 35045 | 93 | acquired | Cluster 73 | wound | Post-intervention | 9/10/2015 | 216 |
| BS3356_R1.fastq | 228C | 440047 | 35045 | 93 | acquired | Cluster 23 | throat | Post-intervention | 9/10/2015 | 216 |
| BS2227_R1.fastq | 237B | 465311 | 35040 | 88 | acquired | Cluster 22 | throat | Pre-intervention | 9/26/2013 | 217 |
| BS2225_R1.fastq | 237A | 465760 | 35040 | 88 | acquired | Cluster 22 | rectal | Pre-intervention | 9/26/2013 | 217 |
| BS2226_R1.fastq | 237C | 465891 | 35043 | 91 | acquired | Cluster 22 | throat | Pre-intervention | 10/3/2013 | 217 |
| BS2234_R1.fastq | 256F | 467996 | 35012 | 60 | acquired | Cluster 54 | throat | Pre-intervention | 10/24/2013 | 219 |
| BS2233_R1.fastq | 256C | 468165 | 35012 | 60 | acquired | Cluster 54 | throat | Pre-intervention | 10/11/2013 | 218 |
| BS2232_R1.fastq | 256D | 468111 | 35012 | 60 | acquired | Cluster 54 | wound | Pre-intervention | 10/11/2013 | 219 |
| BS2231_R1.fastq | 256B | 467618 | 35012 | 60 | acquired | Cluster 54 | rectal | Pre-intervention | 10/11/2013 | 219 |
| BS2230_R1.fastq | 256E | 467540 | 35012 | 60 | acquired | Cluster 54 | sputum | Pre-intervention | 10/21/2013 | 219 |
| BS2229_R1.fastq | 256A | 469654 | 35012 | 60 | acquired | Cluster 54 | sputum | Pre-intervention | 10/7/2013 | 217 |
| BS2235_R1.fastq | 258 | 466193 | 35042 | 90 | acquired | Cluster 22 | throat | Pre-intervention | 10/4/2013 | 217 |
| BS2239_R1.fastq | 294 | 468536 | 35013 | 61 | acquired | Cluster 54 | throat | Pre-intervention | 10/31/2013 | 218 |
| BS3371_R1.fastq | 304A | 495160 | 35011 | 59 | imported | Cluster 20 | sputum | Post-intervention | 11/20/2015 | 219 |
| BS3372_R1.fastq | 304B | 494146 | 35011 | 59 | imported | Cluster 20 | rectal | Post-intervention | 11/20/2015 | 220 |
| BS3374_R1.fastq | 304D | 506259 | 35011 | 59 | imported | Cluster 20 | wound | Post-intervention | 11/24/2015 | 215 |
| BS3376_R1.fastq | 304F | 495306 | 35011 | 59 | imported | Cluster 20 | throat | Post-intervention | 11/24/2015 | 219 |
| BS3373_R1.fastq | 304C | 473605 | 35012 | 60 | imported | Cluster 60 | throat | Post-intervention | 11/20/2015 | 219 |
| BS3375_R1.fastq | 304E | 495028 | 35012 | 60 | imported | Cluster 20 | tissue | Post-intervention | 11/24/2015 | 220 |
| BS3378_R1.fastq | 307 | 468813 | 35012 | 60 | acquired | Cluster 59 | sputum | Post-intervention | 11/24/2015 | 220 |
| BS2240_R1.fastq | 312 | 467752 | 35013 | 61 | imported | Cluster 54 | rectal | Pre-intervention | 4/23/2014 | 218 |
| BS2241_R1.fastq | 318 | 469032 | 35015 | 63 | acquired | Cluster 54 | bronchoalveolar lavage | Pre-intervention | 4/29/2014 | 219 |
| BS3411_R1.fastq | 357 | 457856 | 35011 | 59 | acquired | Cluster 63 | bal | Post-intervention | 1/6/2016 | 219 |
| BS2259_R1.fastq | 364 | 461790 | 35046 | 94 | acquired | Cluster 66 | rectal | Pre-intervention | 5/28/2014 | 217 |
| BS2346_R1.fastq | 393 | 475324 | 35015 | 63 | imported | Cluster 58 | rectal | Pre-intervention | 6/23/2014 | 219 |
| BS2372_R1.fastq | 3_ENV | 469297 | 35014 | 62 | *acquired | Cluster 53 | environment | Pre-intervention | 12/1/2014 | 218 |
| BS2371_R1.fastq | 9_ENV | 466402 | 35014 | 62 | *acquired | Cluster 55 | environment | Pre-intervention | 12/1/2014 | 216 |
| BS2370_R1.fastq | 8B_ENV | 467458 | 35014 | 62 | *acquired | Cluster 56 | environment | Pre-intervention | 12/1/2014 | 219 |
| BS2369_R1.fastq | 8A_ENV | 467089 | 35014 | 62 | *acquired | Cluster 53 | environment | Pre-intervention | 12/1/2014 | 218 |
| BS3414_R1.fastq | 78_ENV | 463029 | 35042 | 90 | *acquired | Cluster 70 | environment | Post-intervention | 1/1/2015 | 218 |

**Supplementary Table 2:** Number of wgSNPs identified in ST357 along with metadata information of strains.

| Strain ID | Isolation ID | Total SNP count | Retained SNPs (With noninformative SNPs) | Retained SNPs with Strict filtring (Without Non-informative SNPs) | Acquired/Imported | MLST PubMLST ST | Sample type | Phase | Sample collection date | Virulence genes numbr |
|---|---|---|---|---|---|---|---|---|---|---|
| BS2360_R1.fastq | 419 | 438141 | 34472 | 13 | acquired | Cluster 7 | rectal | Pre-intervention | 7/15/2014 | 203 |
| BS2354_R1.fastq | 398 | 437578 | 34471 | 12 | acquired | Cluster 6 | rectal | Pre-intervention | 6/25/2014 | 203 |
| BS2352_R1.fastq | 395D | 437062 | 34471 | 12 | imported | Cluster 6 | rectal | Pre-intervention | 6/25/2014 | 202 |
| BS2351_R1.fastq | 395C | 436466 | 34471 | 12 | imported | Cluster 6 | rectal | Pre-intervention | 6/25/2014 | 203 |
| BS2350_R1.fastq | 395F | 436826 | 34471 | 12 | imported | Cluster 6 | throat | Pre-intervention | 6/25/2014 | 202 |
| BS2349_R1.fastq | 395E | 438322 | 34470 | 11 | imported | Cluster 6 | throat | Pre-intervention | 6/25/2014 | 202 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| BS2348_R1.fastq | 395A | 436653 | 34470 | 11 | imported | Cluster 6 | rectal | Pre-intervention | 6/23/2014 | 203 |
| BS2347_R1.fastq | 395B | 439569 | 34470 | 11 | imported | Cluster 6 | throat | Pre-intervention | 6/23/2014 | 203 |
| BS2176_R1.fastq | 46A | 466799 | 34483 | 24 | imported | Cluster 34 | throat | Pre-intervention | 5/10/2013 | 200 |
| BS2170_R1.fastq | 6A | 438169 | 34468 | 9 | acquired | Cluster 6 | rectal | Pre-intervention | 4/10/2013 | 201 |
| BS3246_R1.fastq | 37A | 404253 | 34484 | 25 | acquired | Cluster 39 | rectal | Post-intervention | 2/27/2015 | 204 |
| BS3247_R1.fastq | 37B | 450423 | 34484 | 25 | acquired | Cluster 29 | rectal | Post-intervention | 3/3/2015 | 202 |
| BS3248_R1.fastq | 38A | 433424 | 34484 | 25 | imported | Cluster 1 | sputum | Post-intervention | 2/23/2015 | 203 |
| BS3249_R1.fastq | 38B | 432469 | 34483 | 24 | imported | Cluster 1 | rectal | Post-intervention | 2/23/2015 | 204 |
| BS3250_R1.fastq | 38C | 432478 | 34483 | 24 | imported | Cluster 1 | throat | Post-intervention | 2/23/2015 | 203 |
| BS3256_R1.fastq | 51B | 444890 | 34473 | 14 | acquired | Cluster 3 | throat | Post-intervention | 3/13/2015 | 203 |
| BS3257_R1.fastq | 51C | 441357 | 34473 | 14 | acquired | Cluster 3 | sputum | Post-intervention | 3/16/2015 | 204 |
| BS3266_R1.fastq | 51E | 481394 | 34473 | 14 | acquired | Cluster 3 | throat | Post-intervention | 3/20/2015 | 192 |
| BS3267_R1.fastq | 51F | 438264 | 34473 | 14 | acquired | Cluster 32 | sputum | Post-intervention | 3/31/2015 | 204 |
| BS3269_R1.fastq | 51H | 469271 | 34473 | 14 | acquired | Cluster 3 | wound | Post-intervention | 4/4/2015 | 197 |
| BS3270_R1.fastq | 51I | 437774 | 34473 | 14 | acquired | Cluster 3 | sputum | Post-intervention | 4/4/2015 | 204 |
| BS3271_R1.fastq | 51J | 443738 | 34473 | 14 | acquired | Cluster 3 | throat | Post-intervention | 4/2/2015 | 204 |
| BS3274_R1.fastq | 52A | 443386 | 34471 | 12 | acquired | Cluster 31 | rectal | Post-intervention | 3/13/2015 | 204 |
| BS3276_R1.fastq | 58 | 437071 | 34483 | 24 | acquired | Cluster 1 | rectal | Post-intervention | 3/18/2015 | 203 |
| BS3279_R1.fastq | 66B | 444016 | 34473 | 14 | acquired | Cluster 3 | throat | Post-intervention | 4/6/2015 | 203 |
| BS3280_R1.fastq | 66C | 445691 | 34473 | 14 | acquired | Cluster 3 | throat | Post-intervention | 4/13/2015 | 203 |
| BS3282_R1.fastq | 68B | 437871 | 34483 | 24 | acquired | Cluster 1 | throat | Post-intervention | 3/31/2015 | 205 |
| BS3283_R1.fastq | 68C | 439519 | 34483 | 24 | acquired | Cluster 1 | rectal | Post-intervention | 4/7/2015 | 204 |
| BS3311_R1.fastq | 82A | 439837 | 34473 | 14 | imported | Cluster 3 | throat | Post-intervention | 4/10/2015 | 203 |
| BS3312_R1.fastq | 82B | 440606 | 34473 | 14 | imported | Cluster 3 | throat | Post-intervention | 4/16/2015 | 205 |
| BS3314_R1.fastq | 85B | 448556 | 34480 | 21 | acquired | Cluster 4 | rectal | Post-intervention | 4/20/2015 | 204 |
| BS3316_R1.fastq | 93 | 438286 | 34473 | 14 | imported | Cluster 3 | throat | Post-intervention | 4/20/2015 | 203 |
| BS3317_R1.fastq | 94 | 437098 | 34473 | 14 | acquired | Cluster 33 | tissue | Post-intervention | 5/8/2015 | 204 |
| BS3319_R1.fastq | 104 | 439587 | 34473 | 14 | acquired | Cluster 3 | rectal | Post-intervention | 5/15/2015 | 204 |
| BS3320_R1.fastq | 105 | 437628 | 34473 | 14 | imported | Cluster 3 | throat | Post-intervention | 4/30/2015 | 203 |
| BS3321_R1.fastq | 108 | 432307 | 34485 | 26 | acquired | Cluster 1 | rectal | Post-intervention | 6/11/2015 | 202 |
| BS3322_R1.fastq | 128A | 431781 | 34483 | 24 | acquired | Cluster 1 | throat | Post-intervention | 5/28/2015 | 204 |
| BS3323_R1.fastq | 128B | 434934 | 34472 | 13 | acquired | Cluster 7 | rectal | Post-intervention | 6/3/2015 | 204 |
| BS3324_R1.fastq | 129A | 447524 | 34480 | 21 | acquired | Cluster 4 | bal | Post-intervention | 5/25/2015 | 203 |
| BS3325_R1.fastq | 129B | 446708 | 34480 | 21 | acquired | Cluster 4 | throat | Post-intervention | 5/28/2015 | 204 |
| BS3326_R1.fastq | 129C | 448206 | 34481 | 22 | acquired | Cluster 35 | sputum | Post-intervention | 6/16/2015 | 204 |
| BS3327_R1.fastq | 132A | 435626 | 34472 | 13 | acquired | Cluster 7 | rectal | Post-intervention | 6/11/2015 | 204 |
| BS3328_R1.fastq | 132B | 435845 | 34472 | 13 | acquired | Cluster 7 | rectal | Post-intervention | 6/18/2015 | 204 |
| BS3329_R1.fastq | 161 | 447595 | 34480 | 21 | acquired | Cluster 4 | rectal | Post-intervention | 6/19/2015 | 203 |
| BS3333_R1.fastq | 196 | 497964 | 34477 | 18 | acquired | Cluster 5 | wound | Post-intervention | 7/22/2015 | 199 |
| BS3336_R1.fastq | 203A | 500058 | 34478 | 19 | imported | Cluster 5 | throat | Post-intervention | 7/29/2015 | 199 |
| BS3337_R1.fastq | 203B | 496036 | 34477 | 18 | imported | Cluster 5 | bal | Post-intervention | 7/29/2015 | 198 |
| BS3338_R1.fastq | 203C | 495262 | 34477 | 18 | imported | Cluster 36 | throat | Post-intervention | 7/31/2015 | 199 |
| BS3339_R1.fastq | 203D | 495409 | 34477 | 18 | imported | Cluster 5 | throat | Post-intervention | 8/12/2015 | 199 |
| BS3340_R1.fastq | 203E | 498769 | 34477 | 18 | imported | Cluster 5 | urine | Post-intervention | 9/28/2015 | 200 |
| BS3341_R1.fastq | 206 | 496213 | 34477 | 18 | imported | Cluster 5 | throat | Post-intervention | 7/31/2015 | 199 |
| BS3350_R1.fastq | 222B | 433823 | 34489 | 30 | acquired | Cluster 30 | urine | Post-intervention | 9/18/2015 | 204 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| BS3358_R1.fastq | 236A | 442300 | 34487 | 28 | acquired | Cluster 2 | rectal | Post-intervention | 9/1/2015 | 204 |
| BS3359_R1.fastq | 236B | 440333 | 34487 | 28 | acquired | Cluster 2 | sputum | Post-intervention | 9/7/2015 | 203 |
| BS3360_R1.fastq | 237A | 437425 | 34487 | 28 | imported | Cluster 2 | throat | Post-intervention | 9/1/2015 | 204 |
| BS3362_R1.fastq | 240A | 437234 | 34487 | 28 | acquired | Cluster 2 | rectal | Post-intervention | 7/10/2015 | 203 |
| BS3363_R1.fastq | 240B | 435616 | 34487 | 28 | acquired | Cluster 2 | throat | Post-intervention | 7/10/2015 | 203 |
| BS3364_R1.fastq | 254 | 439401 | 34487 | 28 | acquired | Cluster 2 | throat | Post-intervention | 9/28/2015 | 203 |
| BS3370_R1.fastq | 286 | 433234 | 34487 | 28 | acquired | Cluster 2 | rectal | Post-intervention | 11/3/2015 | 204 |
| BS3379_R1.fastq | 318 | 445596 | 34478 | 19 | imported | Cluster 9 | throat | Post-intervention | 11/30/2015 | 203 |
| BS3380_R1.fastq | 324A | 446254 | 34479 | 20 | imported | Cluster 9 | sputum | Post-intervention | 12/4/2015 | 202 |
| BS3381_R1.fastq | 324B | 446533 | 34479 | 20 | imported | Cluster 9 | throat | Post-intervention | 12/4/2015 | 202 |
| BS3382_R1.fastq | 324C | 447376 | 34479 | 20 | imported | Cluster 9 | rectal | Post-intervention | 12/4/2015 | 203 |
| BS3383_R1.fastq | 324D | 447726 | 34479 | 20 | imported | Cluster 9 | rectal | Post-intervention | 12/8/2015 | 203 |
| BS3384_R1.fastq | 324E | 445393 | 34480 | 21 | imported | Cluster 8 | throat | Post-intervention | 12/8/2015 | 204 |
| BS3385_R1.fastq | 324F | 447331 | 34481 | 22 | imported | Cluster 8 | throat | Post-intervention | 12/21/2015 | 203 |
| BS3386_R1.fastq | 324G | 445733 | 34480 | 21 | imported | Cluster 8 | blood | Post-intervention | 12/28/2015 | 203 |
| BS3387_R1.fastq | 324H | 446881 | 34481 | 22 | imported | Cluster 8 | sputum | Post-intervention | 1/6/2016 | 203 |
| BS3388_R1.fastq | 324I | 444833 | 34480 | 21 | imported | Cluster 38 | blood | Post-intervention | 1/6/2016 | 203 |
| BS3389_R1.fastq | 324J | 446603 | 34480 | 21 | imported | Cluster 37 | blood | Post-intervention | 1/6/2016 | 202 |
| BS3412_R1.fastq | 66_ENV | 434325 | 34469 | 10 | *acquired | Cluster 6 | environment | Post-intervention | 1/1/2015 | 204 |
| BS3413_R1.fastq | 68_ENV | 433868 | 34469 | 10 | *acquired | Cluster 6 | environment | Post-intervention | 1/1/2015 | 204 |

**Supplementary Table 3:** Number of wgSNPs identified in ST823 along with metadata information of strains.

| Strain ID | Isolation ID | Total SNPs | Retained SNPs with non informatv SNPs | Retained SNPs with strict filtering (Without noninformative SNPs) | Acquired/Imported * | MLST PubMLST ST | Sample type | Phase | Sample collection date | Virulence genes numbr |
|---|---|---|---|---|---|---|---|---|---|---|
| BS2379_R1.fastq | 115_ENV | 428343 | 26389 | 7 | *acquired | Cluster 13 | environment | Pre-intervention | 12/1/2014 | 216 |
| BS2368_R1.fastq | 440 | 429344 | 26391 | 9 | acquired | Cluster 12 | throat | Pre-intervention | 8/18/2014 | 217 |
| BS2367_R1.fastq | 412B | 429094 | 26391 | 9 | acquired | Cluster 12 | throat | Pre-intervention | 8/18/2014 | 215 |
| BS2366_R1.fastq | W5C_ENV | 429459 | 26389 | 7 | *acquired | Cluster 12 | environment | Pre-intervention | 8/13/2014 | 216 |
| BS2363_R1.fastq | 421B | 428439 | 26390 | 8 | acquired | Cluster 47 | rectal | Pre-intervention | 7/21/2014 | 215 |
| BS2361_R1.fastq | 420 | 430118 | 26394 | 12 | acquired | Cluster 10 | throat | Pre-intervention | 7/16/2014 | 216 |
| BS2359_R1.fastq | 412A | 429150 | 26390 | 8 | acquired | Cluster 12 | throat | Pre-intervention | 7/15/2014 | 216 |
| BS2358_R1.fastq | 410 | 429825 | 26393 | 11 | acquired | Cluster 44 | sputum | Pre-intervention | 7/14/2014 | 215 |
| BS2355_R1.fastq | 389 | 427641 | 26390 | 8 | acquired | Cluster 50 | sputum | Pre-intervention | 6/27/2014 | 216 |
| BS2345_R1.fastq | 385 | 431586 | 26390 | 8 | acquired | Cluster 12 | sputum | Pre-intervention | 6/17/2014 | 216 |
| BS2344_R1.fastq | 342 | 432345 | 26394 | 12 | acquired | Cluster 10 | sputum | Pre-intervention | 6/11/2014 | 216 |
| BS2343_R1.fastq | 370D | 430315 | 26390 | 8 | imported | Cluster 13 | rectal | Pre-intervention | 6/9/2014 | 216 |
| BS2342_R1.fastq | 370E | 428225 | 26390 | 8 | imported | Cluster 12 | throat | Pre-intervention | 6/9/2014 | 218 |
| BS2341_R1.fastq | 370B | 429233 | 26390 | 8 | imported | Cluster 13 | rectal | Pre-intervention | 6/4/2014 | 216 |
| BS2340_R1.fastq | 370A | 429976 | 26390 | 8 | imported | Cluster 12 | rectal | Pre-intervention | 6/4/2014 | 216 |
| BS2260_R1.fastq | 370C | 428926 | 26390 | 8 | imported | Cluster 13 | sputum | Pre-intervention | 6/4/2014 | 217 |
| BS2258_R1.fastq | 349D | 432649 | 26393 | 11 | acquired | Cluster 10 | urine | Pre-intervention | 5/30/2014 | 217 |
| BS2257_R1.fastq | 349C | 431719 | 26392 | 10 | acquired | Cluster 10 | sputum | Pre-intervention | 5/28/2014 | 216 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| BS2256_R1.fastq | 338 | 426940 | 26391 | 9 | acquired | Cluster 48 | rectal | Pre-intervention | 5/26/2014 | 218 |
| BS2255_R1.fastq | 349B | 429849 | 26393 | 11 | acquired | Cluster 10 | throat | Pre-intervention | 5/23/2014 | 216 |
| BS2254_R1.fastq | 362D | 429643 | 26391 | 9 | imported | Cluster 12 | throat | Pre-intervention | 5/23/2014 | 216 |
| BS2253_R1.fastq | 362C | 429483 | 26389 | 7 | imported | Cluster 12 | throat | Pre-intervention | 5/23/2014 | 217 |
| BS2252_R1.fastq | 362B | 429122 | 26388 | 6 | imported | Cluster 13 | rectal | Pre-intervention | 5/23/2014 | 216 |
| BS2251_R1.fastq | 362A | 427347 | 26388 | 6 | imported | Cluster 13 | rectal | Pre-intervention | 5/23/2014 | 218 |
| BS2250_R1.fastq | 349A | 430496 | 26393 | 11 | acquired | Cluster 10 | throat | Pre-intervention | 5/23/2014 | 216 |
| BS2249_R1.fastq | 351D | 429851 | 26391 | 9 | imported | Cluster 14 | throat | Pre-intervention | 5/18/2014 | 216 |
| BS2248_R1.fastq | 351B | 427602 | 26392 | 10 | imported | Cluster 51 | throat | Pre-intervention | 5/16/2014 | 217 |
| BS2247_R1.fastq | 351C | 429272 | 26391 | 9 | imported | Cluster 14 | rectal | Pre-intervention | 5/18/2014 | 216 |
| BS2246_R1.fastq | 351A | 430959 | 26391 | 9 | imported | Cluster 14 | rectal | Pre-intervention | 5/16/2014 | 216 |
| BS2242_R1.fastq | 317 | 429752 | 26383 | 1 | imported | Cluster 46 | rectal | Pre-intervention | 4/29/2014 | 217 |
| BS2217_R1.fastq | 175 | 1883356 | 26389 | 7 | acquired | Cluster 11 | rectal | Pre-intervention | 8/26/2013 | 217 |
| BS2213_R1.fastq | 162A | 430308 | 26392 | 10 | imported | Cluster 10 | rectal | Pre-intervention | 7/29/2013 | 217 |
| BS2196_R1.fastq | 98G | 430935 | 26386 | 4 | acquired | Cluster 11 | tissue | Pre-intervention | 7/17/2013 | 216 |
| BS2193_R1.fastq | 93B | 434702 | 26387 | 5 | acquired | Cluster 11 | throat | Pre-intervention | 7/1/2013 | 217 |
| BS2192_R1.fastq | 93C | 435463 | 26388 | 6 | acquired | Cluster 11 | sputum | Pre-intervention | 7/2/2013 | 217 |
| BS2172_R1.fastq | 19 | 431261 | 26386 | 4 | acquired | Cluster 11 | sputum | Pre-intervention | 4/16/2013 | 216 |
| BS3240_R1.fastq | 17A | 427248 | 26390 | 8 | acquired | Cluster 12 | rectal | Post-intervention | 2/5/2015 | 217 |
| BS3241_R1.fastq | 17B | 428530 | 26390 | 8 | acquired | Cluster 12 | throat | Post-intervention | 2/5/2015 | 216 |
| BS3245_R1.fastq | 29 | 427838 | 26389 | 7 | imported | Cluster 45 | throat | Post-intervention | 2/12/2015 | 217 |
| BS3344_R1.fastq | 211A | 427449 | 26398 | 16 | acquired | Cluster 15 | throat | Post-intervention | 9/1/2015 | 217 |
| BS3345_R1.fastq | 211B | 430445 | 26398 | 16 | acquired | Cluster 15 | sputum | Post-intervention | 9/7/2015 | 217 |
| BS3365_R1.fastq | 273A | 430937 | 26394 | 12 | acquired | Cluster 42 | rectal | Post-intervention | 10/22/2015 | 217 |
| BS3366_R1.fastq | 273B | 471155 | 26394 | 12 | acquired | Cluster 40 | urine | Post-intervention | 10/26/2015 | 211 |
| BS3367_R1.fastq | 273C | 426941 | 26394 | 12 | acquired | Cluster 41 | rectal | Post-intervention | 10/29/2015 | 217 |
| BS3368_R1.fastq | 273D | 426377 | 26396 | 14 | acquired | Cluster 43 | throat | Post-intervention | 10/29/2015 | 216 |
| BS3409_R1.fastq | 352A | 424923 | 26392 | 10 | imported | Cluster 49 | throat | Post-intervention | 12/29/2015 | 217 |
| BS3410_R1.fastq | 352B | 426037 | 26392 | 10 | imported | Cluster 49 | throat | Post-intervention | 1/8/2016 | 217 |

**Supplementary Table 4:** All informative wgSNPs within ST235 isolates along with their genomic position and annotion.

| # | SNP Position | ST235-specific_ Informative _wgSNPs | Annotation | Non-informative SNPs |
|---|---|---|---|---|
| 1 | 46327 | G -> A | Protein encoding sequence predicted by BioNumerics | 77% G (56/73) |
| 2 | 76816 | C -> T | probable two-component sensor | 99% C (72/73) |
| 3 | 103401 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 4 | 151684 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 5 | 155015 | C -> T | two-component sensor NarX | 99% C (72/73) |
| 6 | 155203 | C -> T | two-component sensor NarX | 56% C (41/73) |
| 7 | 167014 | C -> T | Protein encoding sequence predicted by BioNumerics | 92% C (67/73) |
| 8 | 183428 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 9 | 204162 | C -> G | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 10 | 208909 | C -> T | hypothetical protein | 56% C (41/73) |

| | | | | |
|---|---|---|---|---|
| 11 | 223639 | G -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 12 | 223900 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 13 | 267195 | G -> A | probable transcriptional regulator | 53% G (39/73) |
| 14 | 275498 | C -> T | conserved hypothetical protein | 99% C (72/73) |
| 15 | 276514 | G -> A | conserved hypothetical protein | 99% G (72/73) |
| 16 | 293363 | G -> C | hypothetical protein | 99% G (72/73) |
| 17 | 308939 | A -> C | threonine synthase | 92% A (67/73) |
| 18 | 323856 | C -> T | single-stranded-DNA-specific exonuclease RecJ | 97% C (71/73) |
| 19 | 324179 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 20 | 325800 | G -> A | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 21 | 331708 | C -> T | hypothetical protein | 99% C (72/73) |
| 22 | 344783 | A -> T | probable chemotaxis transducer | 99% A (72/73) |
| 23 | 373711 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 24 | 379005 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 25 | 395505 | A -> G | Protein encoding sequence predicted by BioNumerics | 97% A (71/73) |
| 26 | 446869 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 27 | 463946 | C -> G | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 28 | 468056 | G -> T | probable acyl-CoA dehydrogenase | 99% G (72/73) |
| 29 | 485977 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 30 | 496269 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 31 | 506774 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 32 | 516315 | C -> A | | 53% C (39/73) |
| 33 | 525296 | A -> G | alginate-c5-mannuronan-epimerase AlgG | 97% A (71/73) |
| 34 | 539621 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 35 | 540749 | T -> A | Protein encoding sequence predicted by BioNumerics | 71% T (52/73) |
| 36 | 557106 | G -> T | probable outer membrane protein precursor | 56% G (41/73) |
| 37 | 559060 | G -> T | hypothetical protein | 97% G (71/73) |
| 38 | 560548 | C -> T | Protein encoding sequence predicted by BioNumerics | 90% C (66/73) |
| 39 | 569020 | C -> G | probable ferredoxin | 97% C (71/73) |
| 40 | 601655 | G -> T | conserved hypothetical protein | 53% G (39/73) |
| 41 | 618563 | G -> A | probable ATP-binding component of ABC transporter | 99% G (72/73) |
| 42 | 633377 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 43 | 648137 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 44 | 667824 | G -> A | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 45 | 685253 | G -> T | probable transcriptional regulator | 99% G (72/73) |
| 46 | 687334 | T -> C | Protein encoding sequence predicted by BioNumerics | 56% T (41/73) |
| 47 | 692268 | C -> T | Protein encoding sequence predicted by BioNumerics | 90% C (66/73) |
| 48 | 698702 | G -> C | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 49 | 713669 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 50 | 713932 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 51 | 777251 | C -> T | Protein encoding sequence predicted by BioNumerics | 58% C (42/73) |
| 52 | 778357 | C -> T | Protein encoding sequence predicted by BioNumerics | 92% C (67/73) |
| 53 | 778503 | T -> C | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 54 | 799164 | G -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 55 | 829815 | C -> T | periplasmic tail-specific protease | 92% C (67/73) |
| 56 | 830839 | A -> G | periplasmic tail-specific protease | 99% A (72/73) |

| | | | | |
|---|---|---|---|---|
| 57 | 833848 | T -> A | probable permease of ABC transporter | 97% T (71/73) |
| 58 | 835501 | A -> G | Protein encoding sequence predicted by BioNumerics | 97% A (71/73) |
| 59 | 842774 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 60 | 847481 | C -> G | hypothetical protein | 53% C (39/73) |
| 61 | 878859 | G -> A | probable two-component response regulator | 99% G (72/73) |
| 62 | 910605 | C -> T | hypothetical protein | 71% C (52/73) |
| 63 | 923280 | C -> T | conserved hypothetical protein | 58% C (42/73) |
| 64 | 935706 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 65 | 935887 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 66 | 936081 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 67 | 946326 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 68 | 951728 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 69 | 973293 | C -> A | Protein encoding sequence predicted by BioNumerics | 58% C (42/73) |
| 70 | 974394 | C -> T | Protein encoding sequence predicted by BioNumerics | 92% C (67/73) |
| 71 | 1013689 | C -> T | probable two-component response regulator | 97% C (71/73) |
| 72 | 1028064 | G -> T | amino acid (lysine/arginine/ornithine/histidine/octopine) ABC transporter periplasmic binding protein | 99% G (72/73) |
| 73 | 1070215 | A -> G | Protein encoding sequence predicted by BioNumerics | 97% A (71/73) |
| 74 | 1077843 | G -> C | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 75 | 1080824 | G -> A | hypothetical protein | 97% G (71/73) |
| 76 | 1115963 | A -> C | Protein encoding sequence predicted by BioNumerics | 97% A (71/73) |
| 77 | 1148970 | T -> C | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 78 | 1161904 | T -> C | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 79 | 1168509 | C -> G | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 80 | 1229894 | A -> C | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 81 | 1245071 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 82 | 1252490 | C -> T | Arginine:Pyruvate Transaminase, AruH | 99% C (72/73) |
| 83 | 1262375 | G -> C | topoisomerase IV subunit B | 92% G (67/73) |
| 84 | 1262939 | C -> G | topoisomerase IV subunit B | 90% C (66/73) |
| 85 | 1265258 | T -> G | topoisomerase IV subunit A | 99% T (72/73) |
| 86 | 1265259 | C -> T | topoisomerase IV subunit A | 92% T (67/73) |
| 87 | 1279577 | C -> T | chemotaxis protein MotA | 99% C (72/73) |
| 88 | 1290403 | A -> G | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 89 | 1300480 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 90 | 1341097 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 91 | 1370444 | G -> A | hypothetical protein | 99% G (72/73) |
| 92 | 1414741 | G -> A | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 93 | 1435026 | C -> T | 2,4-dienoyl-CoA reductase FadH2 | 99% C (72/73) |
| 94 | 1456284 | C -> T | Protein encoding sequence predicted by BioNumerics | 77% C (56/73) |
| 95 | 1467535 | G -> T | Protein encoding sequence predicted by BioNumerics | 58% G (42/73) |
| 96 | 1498865 | T -> C | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 97 | 1551102 | C -> T | probable permease of ABC transporter | 99% C (72/73) |
| 98 | 1557603 | C -> T | penicillin-binding protein 1B | 97% C (71/73) |
| 99 | 1604564 | G -> T | cytochrome C-type biogenesis protein CcmH | 58% G (42/73) |
| 100 | 1614174 | T -> A | probable short-chain dehydrogenase | 90% T (66/73) |
| 101 | 1623502 | G -> A | probable two-component sensor | 99% G (72/73) |
| 102 | 1639361 | C -> T | Protein encoding sequence predicted by BioNumerics | 59% C (43/73) |

| | | | | |
|---|---|---|---|---|
| 103 | 1651562 | C -> T | conserved hypothetical protein | 59% C (43/73) |
| 104 | 1652774 | G -> C | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 105 | 1653751 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 106 | 1654242 | C -> T | Protein encoding sequence predicted by BioNumerics | 78% C (57/73) |
| 107 | 1655112 | G -> T | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 108 | 1673272 | T -> C | hypothetical protein | 97% T (71/73) |
| 109 | 1743628 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 110 | 1792998 | T -> G | Protein encoding sequence predicted by BioNumerics | 56% T (41/73) |
| 111 | 1811777 | G -> A | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 112 | 1813154 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 113 | 1828904 | C -> T | probable tonB-dependent receptor | 53% C (39/73) |
| 114 | 1836004 | T -> G | Protein encoding sequence predicted by BioNumerics | 97% T (71/73) |
| 115 | 1839161 | A -> C | probable transcriptional regulator | 99% A (72/73) |
| 116 | 1844060 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 117 | 1848826 | G -> A | probable semialdehyde dehydrogenase | 53% G (39/73) |
| 118 | 1899957 | A -> C | hypothetical protein | 99% A (72/73) |
| 119 | 1907489 | G -> A | conserved hypothetical protein | 97% G (71/73) |
| 120 | 1907885 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 121 | 1938408 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 122 | 1967265 | A -> T | pyocin protein | 95% A (69/73) |
| 123 | 1978237 | G -> A | probable transcriptional regulator | 53% A (39/73) |
| 124 | 1990599 | A -> T | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 125 | 1992225 | C -> G | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 126 | 2002899 | G -> A | hypothetical protein | 56% G (41/73) |
| 127 | 2023174 | T -> A | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 128 | 2053969 | A -> G | hypothetical protein | 53% A (39/73) |
| 129 | 2078725 | A -> G | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 130 | 2091388 | A -> G | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 131 | 2095377 | C -> T | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 132 | 2102797 | C -> A | probable acyl-CoA dehydrogenase | 97% C (71/73) |
| 133 | 2103253 | C -> T | probable acyl-CoA dehydrogenase | 97% C (71/73) |
| 134 | 2105682 | G -> C | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 135 | 2111131 | T -> C | hypothetical protein | 99% T (72/73) |
| 136 | 2112192 | C -> G | probable transcriptional regulator | 99% C (72/73) |
| 137 | 2124402 | G -> A | Transcriptional regulator MvfR | 59% G (43/73) |
| 138 | 2126026 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 139 | 2160091 | C -> T | DNA gyrase subunit A | 95% C (69/73) |
| 140 | 2163765 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 141 | 2205350 | G -> A | heat-shock protein IbpA | 97% G (71/73) |
| 142 | 2225943 | T -> C | Protein encoding sequence predicted by BioNumerics | 56% T (41/73) |
| 143 | 2273653 | T -> C | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 144 | 2290165 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 145 | 2291459 | G -> T | PelG | 56% G (41/73) |
| 146 | 2291990 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 147 | 2301823 | G -> A | conserved hypothetical protein | 97% G (71/73) |
| 148 | 2308069 | C -> A | conserved hypothetical protein | 56% C (41/73) |

| | | | | |
|---|---|---|---|---|
| 149 | 2322937 | C -> T | Protein encoding sequence predicted by BioNumerics | 90% C (66/73) |
| 150 | 2338847 | G -> A | fatty-acid oxidation complex alpha-subunit | 90% G (66/73) |
| 151 | 2357814 | G -> C | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 152 | 2373212 | C -> G | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 153 | 2376811 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 154 | 2396080 | G -> C | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 155 | 2411799 | A -> T | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 156 | 2421034 | C -> T | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 157 | 2434497 | C -> A | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 158 | 2482676 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 159 | 2500083 | T -> C | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 160 | 2510925 | C -> G | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 161 | 2538238 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 162 | 2550013 | G -> A | two-component sensor, CopS | 58% G (42/73) |
| 163 | 2572540 | C -> T | Protein encoding sequence predicted by BioNumerics | 90% C (66/73) |
| 164 | 2593094 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 165 | 2613869 | G -> A | Protein encoding sequence predicted by BioNumerics | 66% G (48/73) |
| 166 | 2670634 | C -> T | probable major facilitator superfamily (MFS) transporter | 99% C (72/73) |
| 167 | 2685477 | C -> T | two-component sensor PfeS | 92% C (67/73) |
| 168 | 2692624 | G -> T | Protein encoding sequence predicted by BioNumerics | 95% G (69/73) |
| 169 | 2705122 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 170 | 2706809 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 171 | 2804212 | T -> C | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 172 | 2807887 | G -> A | two-component sensor, ParS | 53% G (39/73) |
| 173 | 2808187 | G -> A | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 174 | 2876595 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 175 | 2890376 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 176 | 2915053 | T -> C | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 177 | 2916203 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 178 | 2922278 | A -> C | Protein encoding sequence predicted by BioNumerics | 92% A (67/73) |
| 179 | 2923651 | C -> T | probable oxidoreductase | 73% C (53/73) |
| 180 | 2940586 | C -> T | periplasmic beta-glucosidase | 90% C (66/73) |
| 181 | 2950642 | G -> A | transcriptional regulator ExsA | 53% A (39/73) |
| 182 | 3004896 | C -> T | | 99% C (72/73) |
| 183 | 3007332 | T -> G | probable aminotransferase | 53% T (39/73) |
| 184 | 3029635 | G -> A | Protein encoding sequence predicted by BioNumerics | 56% G (41/73) |
| 185 | 3031846 | T -> C | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 186 | 3034174 | A -> G | Protein encoding sequence predicted by BioNumerics | 56% A (41/73) |
| 187 | 3043483 | G -> T | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 188 | 3049217 | C -> A | hypothetical protein | 53% A (39/73) |
| 189 | 3049234 | T -> C | hypothetical protein | 99% T (72/73) |
| 190 | 3123907 | C -> T | Protein encoding sequence predicted by BioNumerics | 90% C (66/73) |
| 191 | 3131558 | C -> G | probable short-chain dehydrogenase | 53% C (39/73) |
| 192 | 3164771 | G -> A | tyrosyl-tRNA synthetase 2 | 53% G (39/73) |
| 193 | 3183769 | G -> A | conserved hypothetical protein | 53% G (39/73) |
| 194 | 3184906 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |

| | | | | |
|---|---|---|---|---|
| 195 | 3187653 | T -> C | hypothetical protein | 97% T (71/73) |
| 196 | 3193124 | C -> A | probable bacteriophage protein | 99% C (72/73) |
| 197 | 3195700 | A -> T | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 198 | 3209744 | G -> T | probable two-component sensor | 90% G (66/73) |
| 199 | 3223330 | T -> C | Protein encoding sequence predicted by BioNumerics | 56% T (41/73) |
| 200 | 3259829 | C -> G | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 201 | 3261020 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 202 | 3261527 | T -> A | conserved hypothetical protein | 90% T (66/73) |
| 203 | 3294475 | A -> C | probable transcriptional regulator | 99% A (72/73) |
| 204 | 3300761 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 205 | 3308254 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 206 | 3314701 | G -> A | probable acyl-CoA dehydrogenase | 97% G (71/73) |
| 207 | 3332951 | G -> A | conserved hypothetical protein | 53% G (39/73) |
| 208 | 3333072 | G -> A | conserved hypothetical protein | 53% A (39/73) |
| 209 | 3395242 | G -> A | S-adenosyl-L-homocysteine hydrolase | 99% G (72/73) |
| 210 | 3395632 | C -> T | S-adenosyl-L-homocysteine hydrolase | 53% C (39/73) |
| 211 | 3399577 | T -> C | probable ATP-dependent RNA helicase | 97% T (71/73) |
| 212 | 3407206 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 213 | 3410054 | C -> G | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 214 | 3429210 | G -> T | twitching motility protein PilG | 73% G (53/73) |
| 215 | 3451150 | G -> A | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 216 | 3464541 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 217 | 3468240 | T -> C | probable aldehyde dehydrogenase | 99% T (72/73) |
| 218 | 3489298 | C -> A | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 219 | 3514341 | C -> T | probable binding protein component of ABC transporter | 97% C (71/73) |
| 220 | 3535259 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 221 | 3544654 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 222 | 3545579 | C -> T | probable DNA polymerase alpha chain | 99% C (72/73) |
| 223 | 3592837 | C -> T | Protein encoding sequence predicted by BioNumerics | 66% C (48/73) |
| 224 | 3598284 | A -> G | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 225 | 3602521 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 226 | 3608075 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 227 | 3636732 | A -> T | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 228 | 3643349 | G -> A | signal peptidase I | 99% G (72/73) |
| 229 | 3646600 | T -> G | pyridoxal phosphate biosynthetic protein PdxJ | 53% G (39/73) |
| 230 | 3651202 | C -> A | probable ATP-dependent protease | 53% C (39/73) |
| 231 | 3656920 | C -> T | proline dehydrogenase PutA | 95% C (69/73) |
| 232 | 3725378 | C -> T | acetate kinase | 99% C (72/73) |
| 233 | 3751276 | C -> T | probable ATP-binding/permease fusion ABC transporter | 53% C (39/73) |
| 234 | 3754677 | G -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 235 | 3760883 | T -> C | aromatic amino acid transport protein AroP2 | 90% T (66/73) |
| 236 | 3768478 | C -> A | transcriptional regulator PhhR | 99% C (72/73) |
| 237 | 3785349 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 238 | 3788289 | C -> G | Protein encoding sequence predicted by BioNumerics | 59% C (43/73) |
| 239 | 3794786 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 240 | 3811401 | T -> G | Protein encoding sequence predicted by BioNumerics | 58% T (42/73) |

| | | | | |
|---|---|---|---|---|
| 241 | 3817401 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 242 | 3819961 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 243 | 3823875 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 244 | 3831639 | T -> G | Protein encoding sequence predicted by BioNumerics | 77% T (56/73) |
| 245 | 3857594 | C -> T | Basic amino acid, basic peptide and imipenem outer membrane porin OprD precursor | 97% C (71/73) |
| 246 | 3857602 | G -> A | Basic amino acid, basic peptide and imipenem outer membrane porin OprD precursor | 99% G (72/73) |
| 247 | 3895688 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 248 | 3904563 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 249 | 3908020 | G -> T | Peptidoglycan associated lipoprotein OprL precursor | 99% G (72/73) |
| 250 | 3911054 | C -> T | | 78% C (57/73) |
| 251 | 3962220 | G -> A | Protein encoding sequence predicted by BioNumerics | 95% G (69/73) |
| 252 | 3974859 | T -> A | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 253 | 3977880 | C -> A | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 254 | 3985941 | T -> A | probable Resistance-Nodulation-Cell Division (RND) efflux transporter | 99% T (72/73) |
| 255 | 3991093 | T -> C | Protein encoding sequence predicted by BioNumerics | 92% T (67/73) |
| 256 | 4019491 | C -> G | UDP-3-O-acyl-N-acetylglucosamine deacetylase | 99% C (72/73) |
| 257 | 4033310 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 258 | 4036795 | A -> G | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 259 | 4066240 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 260 | 4072225 | G -> A | conserved hypothetical protein | 99% G (72/73) |
| 261 | 4076796 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 262 | 4079413 | T -> A | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 263 | 4083085 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 264 | 4104772 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 265 | 4111124 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 266 | 4130059 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 267 | 4145726 | A -> G | Protein encoding sequence predicted by BioNumerics | 53% A (39/73) |
| 268 | 4174707 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 269 | 4237167 | G -> A | Protein encoding sequence predicted by BioNumerics | 77% G (56/73) |
| 270 | 4241518 | G -> A | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 271 | 4244478 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 272 | 4259817 | G -> A | Protein encoding sequence predicted by BioNumerics | 53% A (39/73) |
| 273 | 4286856 | G -> T | Protein encoding sequence predicted by BioNumerics | 95% G (69/73) |
| 274 | 4289222 | G -> T | hypothetical protein | 97% G (71/73) |
| 275 | 4290080 | T -> C | probable cytochrome c | 99% T (72/73) |
| 276 | 4295423 | T -> G | Protein encoding sequence predicted by BioNumerics | 58% T (42/73) |
| 277 | 4295779 | G -> A | probable oxidoreductase | 99% G (72/73) |
| 278 | 4316048 | G -> C | Protein encoding sequence predicted by BioNumerics | 92% G (67/73) |
| 279 | 4349436 | T -> A | Protein encoding sequence predicted by BioNumerics | 58% T (42/73) |
| 280 | 4370358 | G -> A | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 281 | 4374013 | T -> C | Protein encoding sequence predicted by BioNumerics | 93% T (68/73) |
| 282 | 4457577 | C -> T | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 283 | 4471510 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 284 | 4489384 | G -> A | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 285 | 4502475 | G -> A | ATP-binding protease component ClpA | 99% G (72/73) |
| 286 | 4514474 | T -> A | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |

| | | | | |
|---|---|---|---|---|
| 287 | 4549284 | T -> C | hypothetical protein | 90% T (66/73) |
| 288 | 4573223 | G -> C | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 289 | 4583547 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 290 | 4602220 | C -> A | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 291 | 4606648 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 292 | 4623188 | C -> T | probable aldehyde dehydrogenase | 90% C (66/73) |
| 293 | 4649090 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 294 | 4689026 | C -> T | probable iron-sulfur protein | 58% C (42/73) |
| 295 | 4705878 | G -> A | probable toxin transporter | 97% G (71/73) |
| 296 | 4708578 | G -> A | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 297 | 4716634 | G -> A | probable hydrolase | 53% A (39/73) |
| 298 | 4721652 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (66/73) |
| 299 | 4735142 | A -> T | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 300 | 4745746 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 301 | 4758000 | G -> A | probable transcriptional regulator | 53% A (39/73) |
| 302 | 4785106 | C -> G | probable outer membrane protein precursor | 90% C (66/73) |
| 303 | 4785442 | G -> A | probable outer membrane protein precursor | 90% G (66/73) |
| 304 | 4793086 | G -> A | tryptophan synthase alpha chain | 92% G (67/73) |
| 305 | 4840699 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 306 | 4842923 | C -> A | Protein encoding sequence predicted by BioNumerics | 53% A (39/73) |
| 307 | 4859377 | G -> A | hypothetical protein | 97% G (71/73) |
| 308 | 4904449 | C -> A | Protein encoding sequence predicted by BioNumerics | 92% C (67/73) |
| 309 | 4919357 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 310 | 4920523 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 311 | 4922252 | G -> T | Protein encoding sequence predicted by BioNumerics | 53% G (39/73) |
| 312 | 4952190 | C -> T | Protein encoding sequence predicted by BioNumerics | 71% C (52/73) |
| 313 | 4959430 | T -> C | glycosyltransferase WbpZ | 99% T (72/73) |
| 314 | 4965290 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 315 | 4975071 | C -> A | probable transcriptional regulator | 53% A (39/73) |
| 316 | 4987350 | G -> A | probable transcriptional regulator | 97% G (71/73) |
| 317 | 5035045 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 318 | 5072781 | G -> T | hypothetical protein | 53% G (39/73) |
| 319 | 5079548 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 320 | 5096709 | T -> C | probable aldehyde dehydrogenase | 97% T (71/73) |
| 321 | 5103216 | C -> T | fimbrial subunit CupA4 | 99% C (72/73) |
| 322 | 5122689 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 323 | 5127186 | G -> A | Protein encoding sequence predicted by BioNumerics | 58% G (42/73) |
| 324 | 5128210 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 325 | 5135250 | T -> C | probable alcohol dehydrogenase (Zn-dependent) | 95% T (69/73) |
| 326 | 5140170 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 327 | 5144577 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 328 | 5150318 | C -> T | hypothetical protein | 99% C (72/73) |
| 329 | 5156260 | G -> A | probable sensor/response regulator hybrid | 97% G (71/73) |
| 330 | 5180828 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 331 | 5183096 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 332 | 5219670 | C -> T | | 97% C (71/73) |

| | | | | |
|---|---|---|---|---|
| 333 | 5237789 | C -> T | hypothetical protein | 99% C (72/73) |
| 334 | 5244778 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 335 | 5260534 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 336 | 5261063 | C -> T | histidine porin OpdC | 58% C (42/73) |
| 337 | 5265936 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 338 | 5278930 | G -> A | Protein encoding sequence predicted by BioNumerics | 58% G (42/73) |
| 339 | 5281273 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 340 | 5292436 | C -> T | Protein encoding sequence predicted by BioNumerics | 66% C (48/73) |
| 341 | 5295004 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 342 | 5306205 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 343 | 5309436 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 344 | 5316193 | G -> C | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 345 | 5326737 | T -> C | Protein encoding sequence predicted by BioNumerics | 90% C (66/73) |
| 346 | 5336235 | G -> A | probable transcriptional regulator | 99% G (72/73) |
| 347 | 5338926 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% C (39/73) |
| 348 | 5393479 | C -> T | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 349 | 5397505 | G -> A | | 99% G (72/73) |
| 350 | 5439723 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (72/73) |
| 351 | 5480375 | G -> A | Protein encoding sequence predicted by BioNumerics | 58% G (42/73) |
| 352 | 5490072 | C -> A | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 353 | 5603909 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 354 | 5607912 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 355 | 5616071 | G -> A | ferric-mycobactin receptor, FemA | 53% A (39/73) |
| 356 | 5620315 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 357 | 5633388 | G -> A | hypothetical protein | 99% G (72/73) |
| 358 | 5633678 | G -> A | hypothetical protein | 97% G (71/73) |
| 359 | 5662767 | C -> G | ribose transport protein RbsA | 99% C (72/73) |
| 360 | 5697631 | C -> T | hypothetical protein | 53% C (39/73) |
| 361 | 5700030 | A -> G | cytochrome c550 | 99% A (72/73) |
| 362 | 5700253 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 363 | 5706567 | C -> T | Protein encoding sequence predicted by BioNumerics | 97% C (71/73) |
| 364 | 5709617 | C -> T | hypothetical protein | 77% C (56/73) |
| 365 | 5721230 | G -> A | Protein encoding sequence predicted by BioNumerics | 66% G (48/73) |
| 366 | 5736541 | G -> A | paerucumarin biosynthesis protein PvcC | 58% G (42/73) |
| 367 | 5801854 | A -> T | Protein encoding sequence predicted by BioNumerics | 95% A (69/73) |
| 368 | 5807584 | T -> A, G | | 53% A (39/73) |
| 369 | 5874173 | C -> T | probable TonB-dependent receptor | 53% C (39/73) |
| 370 | 5889363 | C -> T | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 371 | 5928898 | T -> C | hypothetical protein | 53% T (39/73) |
| 372 | 5933535 | C -> T | hypothetical protein | 56% C (41/73) |
| 373 | 5940449 | A -> C | Protein encoding sequence predicted by BioNumerics | 56% A (41/73) |
| 374 | 5981666 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 375 | 5997016 | C -> A | Protein encoding sequence predicted by BioNumerics | 56% C (41/73) |
| 376 | 5997592 | C -> T | Protein encoding sequence predicted by BioNumerics | 95% C (69/73) |
| 377 | 6007368 | A -> G | Protein encoding sequence predicted by BioNumerics | 99% A (72/73) |
| 378 | 6017170 | C -> T | elongation factor G | 99% C (72/73) |

| | | | | |
|---|---|---|---|---|
| 379 | 6057123 | T -> A | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 380 | 6068184 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 381 | 6075633 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 382 | 6076399 | A -> G | probable transcriptional regulator | 56% A (41/73) |
| 383 | 6079620 | C -> G | conserved hypothetical protein | 56% C (41/73) |
| 384 | 6091816 | C -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 385 | 6093040 | G -> A | putative isovaleryl-CoA dehydrogenase | 99% G (72/73) |
| 386 | 6102951 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 387 | 6111857 | C -> T | DhcB, dehydrocarnitine CoA transferase, subunit B | 68% C (50/73) |
| 388 | 6122137 | C -> T | Protein encoding sequence predicted by BioNumerics | 92% C (67/73) |
| 389 | 6161082 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (72/73) |
| 390 | 6177766 | G -> A | exonuclease SbcD | 53% A (39/73) |
| 391 | 6200140 | T -> C | two-component sensor PprA | 97% T (71/73) |
| 392 | 6239120 | C -> T | probable major facilitator superfamily (MFS) transporter | 99% C (72/73) |
| 393 | 6247301 | G -> C | Protein encoding sequence predicted by BioNumerics | 97% G (71/73) |
| 394 | 6250826 | G -> T | Protein encoding sequence predicted by BioNumerics | 53% T (39/73) |
| 395 | 6256972 | G -> A | hypothetical protein | 95% G (69/73) |
| 396 | 6272087 | T -> C | Protein encoding sequence predicted by BioNumerics | 97% T (71/73) |
| 397 | 6272134 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (72/73) |
| 398 | 6272885 | G -> A | L-2,4-diaminobutyrate:2-ketoglutarate 4-aminotransferase, PvdH | 90% G (66/73) |

**Supplementary Table 5:** All informative wgSNPs within ST357 isolates along with their genomic position and annotion.

| # | SNP Position | ST357-specific_ informative_wgSNPs | Annotaion | Non-informative SNPs |
|---|---|---|---|---|
| 1 | 8881 | C -> T | Protein encoding sequence predicted by BioNumerics | 93% C (67/72) |
| 2 | 54864 | G -> C | Protein encoding sequence predicted by BioNumerics | 81% G (58/72) |
| 3 | 86136 | C -> T | probable permease of ABC taurine transporter | 99% C (71/72) |
| 4 | 86158 | G -> T | probable permease of ABC taurine transporter | 81% G (58/72) |
| 5 | 88557 | C -> T | Protein encoding sequence predicted by BioNumerics | 92% C (66/72) |
| 6 | 274331 | C -> G | probable aromatic amino acid transporter (aroP2) | 99% C (71/72) |
| 7 | 277484 | G -> A | phosphoribosylformylglycinamidine synthase (purl) | 65% A (47/72) |
| 8 | 285092 | C -> T | N-Acetyl-D-Glucosamine phosphotransferase system transporter (PA3760) | 90% C (65/72) |
| 9 | 324358 | C -> T | Protein encoding sequence predicted by BioNumerics | 99% C (71/72) |
| 10 | 347621 | G -> T | probable chemotaxis sensor/effector fusion protein | 75% G (54/72) |
| 11 | 422118 | T -> A | Protein encoding sequence predicted by BioNumerics | 90% T (65/72) |
| 12 | 442552 | G -> A | hypothetical protein | 81% G (58/72) |
| 13 | 533317 | A -> G | | 88% A (63/72) |
| 14 | 537314 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (71/72) |
| 15 | 581522 | G -> A | Protein encoding sequence predicted by BioNumerics | 75% G (54/72) |
| 16 | 587150 | C -> G | probable malic enzyme (mdh) | 65% G (47/72) |
| 17 | 638412 | G -> T | Protein encoding sequence predicted by BioNumerics | 65% T (47/72) |
| 18 | 942202 | T -> A | probable ATP-binding/permease fusion ABC transporter | 75% T (54/72) |
| 19 | 962972 | T -> G | probable secretion pathway ATPase | 97% T (70/72) |
| 20 | 994923 | G -> A | probable oxidoreductase | 81% G (58/72) |

| | | | | |
|---|---|---|---|---|
| 21 | 1109585 | C -> T | conserved hypothetical protein | 99% C (71/72) |
| 22 | 1157825 | T -> G | Protein encoding sequence predicted by BioNumerics | 69% T (50/72) |
| 23 | 1159859 | A -> C | Protein encoding sequence predicted by BioNumerics | 99% A (71/72) |
| 24 | 1160329 | G -> T | type 4 fimbrial biogenesis protein PilO | 75% G (54/72) |
| 25 | 1177023 | G -> A | Protein encoding sequence predicted by BioNumerics | 79% G (57/72) |
| 26 | 1197662 | T -> G | conserved hypothetical protein | 90% T (65/72) |
| 27 | 1203653 | C -> A | glutamate-ammonia-ligase adenylyltransferase | 93% C (67/72) |
| 28 | 1273875 | G -> T | Protein encoding sequence predicted by BioNumerics | 75% G (54/72) |
| 29 | 1408777 | C -> T | Protein encoding sequence predicted by BioNumerics | 85% C (61/72) |
| 30 | 1486776 | G -> A | DNA repair protein RecN | 79% G (57/72) |
| 31 | 1624337 | A -> C | Protein encoding sequence predicted by BioNumerics | 90% A (65/72) |
| 32 | 1644287 | G -> A | probable two-component sensor | 99% A (71/72) |
| 33 | 1680012 | C -> T | Protein encoding sequence predicted by BioNumerics | 81% C (58/72) |
| 34 | 1691207 | G -> C | probable pyruvate carboxylase | 81% G (58/72) |
| 35 | 1747575 | C -> A | Protein encoding sequence predicted by BioNumerics | 85% C (61/72) |
| 36 | 1784380 | G -> A | Protein encoding sequence predicted by BioNumerics | 97% G (70/72) |
| 37 | 1878577 | C -> A | hypothetical protein | 75% C (54/72) |
| 38 | 1912394 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (65/72) |
| 39 | 1951557 | C -> T | Protein encoding sequence predicted by BioNumerics | 75% C (54/72) |
| 40 | 1953742 | A -> C | rRNA methyltransferase | 92% A (66/72) |
| 41 | 1990381 | A -> G | Protein encoding sequence predicted by BioNumerics | 89% A (64/72) |
| 42 | 2007565 | C -> A | conserved hypothetical protein | 93% C (67/72) |
| 43 | 2046754 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (71/72) |
| 44 | 2100548 | C -> T | 2-Nitropropane Dioxygenase | 75% C (54/72) |
| 45 | 2498002 | C -> T | lipase modulator protein | 65% T (47/72) |
| 46 | 2807068 | G -> A | trigger factor | 94% G (68/72) |
| 47 | 2807859 | C -> G | two-component sensor, ParS | 92% C (66/72) |
| 48 | 2866716 | G -> A | hypothetical protein | 85% G (61/72) |
| 49 | 2871796 | C -> G | aconitate hydratase 2 | 75% C (54/72) |
| 50 | 2903551 | T -> C | Protein encoding sequence predicted by BioNumerics | 65% C (47/72) |
| 51 | 2950278 | G -> T | transcriptional regulator ExsA | 99% G (71/72) |
| 52 | 2958965 | C -> A | Protein encoding sequence predicted by BioNumerics | 99% C (71/72) |
| 53 | 3025078 | T -> A | Protein encoding sequence predicted by BioNumerics | 99% T (71/72) |
| 54 | 3077679 | T -> A | Protein encoding sequence predicted by BioNumerics | 85% T (61/72) |
| 55 | 3105952 | A -> G | Protein encoding sequence predicted by BioNumerics | 81% A (58/72) |
| 56 | 3120476 | G -> A | Protein encoding sequence predicted by BioNumerics | 94% G (68/72) |
| 57 | 3221380 | A -> T | Protein encoding sequence predicted by BioNumerics | 90% A (65/72) |
| 58 | 3239702 | A -> T | Protein encoding sequence predicted by BioNumerics | 75% A (54/72) |
| 59 | 3321398 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (71/72) |
| 60 | 3481273 | C -> T | dihydroxy-acid dehydratase | 93% C (67/72) |
| 61 | 3507162 | C -> A | Protein encoding sequence predicted by BioNumerics | 90% C (65/72) |
| 62 | 3531945 | G -> A | probable transcriptional regulator | 79% G (57/72) |
| 63 | 3550584 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (71/72) |
| 64 | 3593783 | C -> T | probable major facilitator superfamily (MFS) transporter | 85% C (61/72) |
| 65 | 3639711 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (71/72) |
| 66 | 3705544 | C -> A | Protein encoding sequence predicted by BioNumerics | 85% C (61/72) |

| | | | | |
|---|---|---|---|---|
| 67 | 3744759 | A -> G | Protein encoding sequence predicted by BioNumerics | 65% G (47/72) |
| 68 | 3752954 | C -> A | Protein encoding sequence predicted by BioNumerics | 81% C (58/72) |
| 69 | 3754984 | G -> A | Protein encoding sequence predicted by BioNumerics | 85% G (61/72) |
| 70 | 3838501 | A -> G | GTP pyrophosphokinase (relA) | 85% A (61/72) |
| 71 | 3857898 | C -> T | Basic amino acid, basic peptide and imipenem outer membrane porin OprD precursor | 99% C (71/72) |
| 72 | 3953805 | G -> T | probable hydrolase | 93% G (67/72) |
| 73 | 4033310 | G -> A | Protein encoding sequence predicted by BioNumerics | 94% G (68/72) |
| 74 | 4061815 | G -> C | Protein encoding sequence predicted by BioNumerics | 85% G (61/72) |
| 75 | 4072416 | C -> T | probable ATP-binding component of ABC transporter | 90% C (65/72) |
| 76 | 4120385 | A -> C | Protein encoding sequence predicted by BioNumerics | 93% A (67/72) |
| 77 | 4145692 | G -> A | Protein encoding sequence predicted by BioNumerics | 65% A (47/72) |
| 78 | 4319886 | C -> A | Protein encoding sequence predicted by BioNumerics | 96% C (69/72) |
| 79 | 4332627 | G -> T | | 85% G (61/72) |
| 80 | 4391947 | T -> G | hypothetical protein | 86% T (62/72) |
| 81 | 4392547 | C -> A | Protein encoding sequence predicted by BioNumerics | 85% C (61/72) |
| 82 | 4479451 | C -> T | hypothetical protein | 92% C (66/72) |
| 83 | 4556323 | G -> A | Protein encoding sequence predicted by BioNumerics | 99% G (71/72) |
| 84 | 4600943 | G -> A | conserved hypothetical protein | 99% G (71/72) |
| 85 | 4624783 | C -> T | probable transcriptional regulator | 99% C (71/72) |
| 86 | 4709257 | C -> T | probable outer membrane protein precursor | 90% C (65/72) |
| 87 | 4817906 | G -> A | Protein encoding sequence predicted by BioNumerics | 75% G (54/72) |
| 88 | 4849885 | T -> G | Protein encoding sequence predicted by BioNumerics | 93% T (67/72) |
| 89 | 4912287 | C -> T | | 85% C (61/72) |
| 90 | 4915615 | G -> A | hypothetical protein | 99% G (71/72) |
| 91 | 4966735 | C -> T | Protein encoding sequence predicted by BioNumerics | 94% C (68/72) |
| 92 | 4975543 | T -> G | Protein encoding sequence predicted by BioNumerics | 99% T (71/72) |
| 93 | 4979139 | G -> T | probable transcarboxylase subunit | 99% G (71/72) |
| 94 | 5052925 | A -> C | Protein encoding sequence predicted by BioNumerics | 99% A (71/72) |
| 95 | 5176542 | G -> T | Protein encoding sequence predicted by BioNumerics | 85% G (61/72) |
| 96 | 5223387 | A -> C | Protein encoding sequence predicted by BioNumerics | 82% A (59/72) |
| 97 | 5250949 | G -> C | Protein encoding sequence predicted by BioNumerics | 75% G (54/72) |
| 98 | 5396238 | A -> G | Protein encoding sequence predicted by BioNumerics | 85% A (61/72) |
| 99 | 5422044 | C -> T | Protein encoding sequence predicted by BioNumerics | 75% C (54/72) |
| 100 | 5427561 | G -> T | Protein encoding sequence predicted by BioNumerics | 99% G (71/72) |
| 101 | 5444428 | G -> T | Protein encoding sequence predicted by BioNumerics | 65% T (47/72) |
| 102 | 5466897 | C -> T | gluconate permease (gnuT) | 75% C (54/72) |
| 103 | 5587591 | G -> A | Protein encoding sequence predicted by BioNumerics | 90% G (65/72) |
| 104 | 5699927 | A -> G | cytochrome c550 (exaB) | 94% A (68/72) |
| 105 | 5839714 | C -> T | hypothetical protein | 85% C (61/72) |
| 106 | 5950632 | A -> G | Protein encoding sequence predicted by BioNumerics | 99% A (71/72) |
| 107 | 5976296 | C -> T | Protein encoding sequence predicted by BioNumerics | 93% C (67/72) |
| 108 | 6024859 | C -> T | 30S ribosomal protein S17 (rpsQ) | 88% C (63/72) |
| 109 | 6061009 | C -> T | Fe(III)-pyochelin outer membrane receptor precursor (fptA) | 65% T (47/72) |
| 110 | 6066230 | C -> A | Protein encoding sequence predicted by BioNumerics | 85% C (61/72) |
| 111 | 6119026 | C -> G | ErcS | 93% C (67/72) |

**Supplementary Table 6:** All informative wgSNPs within ST823 isolates along with their genomic position and annotion.

| # | SNP Position | ST823-specific_ informative_wgSNPs | Annotation | Non-informative SNPs |
|---|---|---|---|---|
| 1 | 3262 | C -> T | UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase (mpl gene) | 91% C (43/47) |
| 2 | 140503 | C -> T | conserved hypothetical protein | 98% C (46/47) |
| 3 | 179295 | C -> T | Protein encoding sequence predicted by BioNumerics | 98% C (46/47) |
| 4 | 238311 | C -> T | conserved hypothetical protein | 98% C (46/47) |
| 5 | 260110 | C -> T | Protein encoding sequence predicted by BioNumerics | 91% C (43/47) |
| 6 | 293935 | C -> T | Protein encoding sequence predicted by BioNumerics | 98% C (46/47) |
| 7 | 329965 | A -> G | Protein encoding sequence predicted by BioNumerics | 68% A (32/47) |
| 8 | 352043 | C -> T | Protein encoding sequence predicted by BioNumerics | 98% C (46/47) |
| 9 | 474090 | G -> T | Protein encoding sequence predicted by BioNumerics | 91% G (43/47) |
| 10 | 492449 | G -> A | Protein encoding sequence predicted by BioNumerics | 91% G (43/47) |
| 11 | 640270 | G -> A | conserved hypothetical protein | 96% G (45/47) |
| 12 | 812458 | C -> T | Protein encoding sequence predicted by BioNumerics | 83% C (39/47) |
| 13 | 1595002 | T -> A | Protein encoding sequence predicted by BioNumerics | 83% T (39/47) |
| 14 | 1653618 | G -> T | Protein encoding sequence predicted by BioNumerics | 62% T (29/47) |
| 15 | 1653760 | G -> A | Protein encoding sequence predicted by BioNumerics | 83% G (39/47) |
| 16 | 1797565 | G -> T | Protein encoding sequence predicted by BioNumerics | 96% G (45/47) |
| 17 | 1816213 | T -> G | Protein encoding sequence predicted by BioNumerics | 91% T (43/47) |
| 18 | 1867219 | T -> A | hypothetical protein | 89% T (42/47) |
| 19 | 1874535 | C -> T | Protein encoding sequence predicted by BioNumerics | 98% C (46/47) |
| 20 | 1903856 | G -> A | Protein encoding sequence predicted by BioNumerics | 91% G (43/47) |
| 21 | 1920917 | G -> T | Protein encoding sequence predicted by BioNumerics | 81% G (38/47) |
| 22 | 1950618 | G -> T | Protein encoding sequence predicted by BioNumerics | 89% G (42/47) |
| 23 | 1990561 | C -> A, T | Protein encoding sequence predicted by BioNumerics | 96% A (45/47) |
| 24 | 2023500 | G -> A | Protein encoding sequence predicted by BioNumerics | 98% G (46/47) |
| 25 | 2076496 | G -> T | Protein encoding sequence predicted by BioNumerics | 96% G (45/47) |
| 26 | 2087747 | G -> A | Protein encoding sequence predicted by BioNumerics | 64% G (30/47) |
| 27 | 2170248 | C -> T | Protein encoding sequence predicted by BioNumerics | 91% C (43/47) |
| 28 | 2188316 | T -> C | Protein encoding sequence predicted by BioNumerics | 91% T (43/47) |
| 29 | 2207630 | G -> C | Protein encoding sequence predicted by BioNumerics | 98% G (46/47) |
| 30 | 2311875 | C -> T | Protein encoding sequence predicted by BioNumerics | 83% C (39/47) |
| 31 | 2341769 | A -> G | DNA topoisomerase I | 94% A (44/47) |
| 32 | 2912817 | C -> G | Protein encoding sequence predicted by BioNumerics | 64% C (30/47) |
| 33 | 3362063 | T -> A | probable ClpA/B protease ATP binding subunit | 83% T (39/47) |
| 34 | 3509275 | C -> T | hypothetical protein | 64% C (30/47) |
| 35 | 3594828 | G -> A | Protein encoding sequence predicted by BioNumerics | 64% G (30/47) |
| 36 | 3808176 | C -> G | Protein encoding sequence predicted by BioNumerics | 96% C (45/47) |
| 37 | 3828139 | C -> A | sensor/response regulator hybrid | 87% C (41/47) |
| 38 | 3904922 | G -> T | Protein encoding sequence predicted by BioNumerics | 94% G (44/47) |
| 39 | 3919236 | G -> A | Protein encoding sequence predicted by BioNumerics | 98% G (46/47) |
| 40 | 4627927 | A -> C | Protein encoding sequence predicted by BioNumerics | 62% C (29/47) |
| 41 | 4704175 | G -> C | Protein encoding sequence predicted by BioNumerics | 98% G (46/47) |
| 42 | 4810751 | G -> A | conserved hypothetical protein | 98% G (46/47) |

| | | | | |
|---|---|---|---|---|
| 43 | 4862083 | C -> G | dihydroorotase (DHODH) | 96% C (45/47) |
| 44 | 4887501 | C -> A | Protein encoding sequence predicted by BioNumerics | 96% C (45/47) |
| 45 | 5049615 | C -> T | Protein encoding sequence predicted by BioNumerics | 96% C (45/47) |
| 46 | 5065005 | G -> A | hypothetical protein | 98% G (46/47) |
| 47 | 5180230 | C -> T | hypothetical protein | 81% C (38/47) |
| 48 | 5200128 | G -> C | probable carbonic anhydrase | 96% G (45/47) |
| 49 | 5354684 | A -> T | Protein encoding sequence predicted by BioNumerics | 98% A (46/47) |
| 50 | 5438644 | G -> T | Protein encoding sequence predicted by BioNumerics | 96% G (45/47) |
| 51 | 5518279 | T -> C | hypothetical protein | 62% C (29/47) |
| 52 | 5529241 | C -> T | ATP-dependent DNA helicase RecG | 98% C (46/47) |
| 53 | 5625095 | T -> A | Protein encoding sequence predicted by BioNumerics | 91% T (43/47) |
| 54 | 5646282 | C -> T | Protein encoding sequence predicted by BioNumerics | 68% C (32/47) |
| 55 | 5888963 | C -> T | Protein encoding sequence predicted by BioNumerics | 91% C (43/47) |
| 56 | 5962111 | C -> T | PvdL | 51% T (24/47) |
| 57 | 6026688 | C -> T | Protein encoding sequence predicted by BioNumerics | 91% C (43/47) |
| 58 | 6026753 | C -> A | Protein encoding sequence predicted by BioNumerics | 96% C (45/47) |
| 59 | 6124007 | G -> A | Protein encoding sequence predicted by BioNumerics | 98% G (46/47) |
| 60 | 6224114 | A -> G | Protein encoding sequence predicted by BioNumerics | 96% A (45/47) |
| 61 | 6283222 | C -> T | probable transcriptional regulator | 98% C (46/47) |

**References**


ASHTON, P. M., NAIR, S., PETERS, T. M., BALE, J. A., POWELL, D. G., PAINSET, A., TEWOLDE, R., SCHAEFER, U., JENKINS, C., DALLMAN, T. J., DE PINNA, E. M., GRANT, K. A. & SALMONELLA WHOLE GENOME SEQUENCING IMPLEMENTATION, G. 2016b. Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ,* 4**,** e1752.

BAKKER, H. C., SWITT, A. I., CUMMINGS, C. A., HOELZER, K., DEGORICIJA, L., RODRIGUEZ-RIVERA, L. D., WRIGHT, E. M., FANG, R., DAVIS, M., ROOT, T., SCHOONMAKER-BOPP, D., MUSSER, K. A., VILLAMIL, E., WAECHTER, H., KORNSTEIN, L., FURTADO, M. R. & WIEDMANN, M. 2011. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common Salmonella enterica subsp. enterica serovar Montevideo pulsed-field gel electrophoresis type. *Appl Environ Microbiol,* 77**,** 8648-55.

BEDARD, E., PREVOST, M. & DEZIEL, E. 2016. Pseudomonas aeruginosa in premise plumbing of large buildings. *Microbiologyopen,* 5**,** 937-956.

BLANC, D. S., MAGALHAES, B., KOENIG, I., SENN, L. & GRANDBASTIEN, B. 2020. Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumerics(TM)) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa. *Front Microbiol,* 11**,** 1729.

CABOT, G., ZAMORANO, L., MOYA, B., JUAN, C., NAVAS, A., BLAZQUEZ, J. & OLIVER, A. 2016. Evolution of Pseudomonas aeruginosa Antimicrobial Resistance and Fitness under Low and High Mutation Rates. *Antimicrob Agents Chemother,* 60**,** 1767-78.

CDC 2019. ANTIBIOTIC RESISTANCE THREATS IN THE UNITED STATES 2019. Atlanta, GA:: U.S. Department of Health and Human Services, CDC; 2019.

CHEN, L., ZHENG, D., LIU, B., YANG, J. & JIN, Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res,* 44**,** D694-7.

CODY, A. J., MCCARTHY, N. D., RENSBURG, M. J. V., ISINKAYE, T., BENTLEY, S. D., PARKHILL, J., DINGLE, K. E., BOWLER, I. C. J. W., JOLLEY, K. A. & MAIDEN, M. C. J. 2013. Real-Time Genomic Epidemiological Evaluation of Human Campylobacter Isolates by Use of Whole-Genome Multilocus Sequence Typing. *Journal of Clinical Microbiology* 51**,** 2526–2534.

COWLEY, L. A., PETERSEN, F. C., JUNGES, R., JIMSON, D. J. M., MORRISON, D. A. & HANAGE, W. P. 2018. Evolution via recombination: Cell-to-cell contact facilitates larger recombination events in Streptococcus pneumoniae. *PLoS Genet,* 14**,** e1007410.

DE OLIVEIRA, D., FORDE, B., KIDD, T., HARRIS, P., SCHEMBRI, M., BEATSON, S., PATERSON, D. & WALKER, M. 2020. Antimicrobial Resistance in ESKAPE Pathogens. *Clin Microbiol Rev* 33**,** e00181-19.

DORING, G., PARAMESWARAN, I. G. & MURPHY, T. F. 2011. Differential adaptation of microbial pathogens to airways of patients with cystic fibrosis and chronic obstructive pulmonary disease. *FEMS Microbiol Rev,* 35**,** 124-46.

FUJII, A., SEKI, M., HIGASHIGUCHI, M., TACHIBANA, I., KUMANOGOH, A. & TOMONO, K. 2014. Community-acquired, hospital-acquired, and healthcare-associated pneumonia caused by Pseudomonas aeruginosa. *Respiratory Medicine Case Reports,* 12**,** 30-33.

GAD, S. C. 2014. Epidemiology. 433-437.

GATEAU, C., DEBOSCKER, S., COUTURIER, J., VOGEL, T., SCHMITT, E., MULLER, J., MENARD, C., TURCAN, B., ZAIDI, R. S., YOUSSOUF, A., LAVIGNE, T. &

BARBUT, F. 2019b. Local outbreak of Clostridioides difficile PCR-Ribotype 018 investigated by multi locus variable number tandem repeat analysis, whole genome multi locus sequence typing and core genome single nucleotide polymorphism typing. *Anaerobe,* 60**,** 102087.

HALACHEV, M. R., CHAN, J. Z., CONSTANTINIDOU, C. I., CUMLEY, N., BRADLEY, C., SMITH-BANKS, M., OPPENHEIM, B. & PALLEN, M. J. 2014a. Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant Acinetobacter baumannii in Birmingham, England. *Genome Med,* 6**,** 70.

HEALTH, U. D. O. & SERVICES, H. 2019. CDC. Antibiotic resistance threats in the United States, 2019. CDC Atlanta: Atlanta, GA, USA.

INNS, T., HAWKER, J., ELSON, R., NEAL, K., ADAK, G. K., LANE, C., PETERS, T., DALLMAN, T., CHATT, C., MCFARLAND, N., CROOK, P., BISHOP, T. & TEAM, P. C. O. B. O. T. O. C. 2015a. A multi-country Salmonella Enteritidis phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom,May to September 2014. *Euro Surveill*

JACOBS, D. M., OCHS-BALCOM, H. M., NOYES, K., ZHAO, J., LEUNG, W. Y., PU, C. Y., MURPHY, T. F. & SETHI, S. 2020. Impact of Pseudomonas aeruginosa Isolation on Mortality and Outcomes in an Outpatient Chronic Obstructive Pulmonary Disease Cohort. *Open Forum Infect Dis,* 7**,** ofz546.

JOENSEN, K. G., SCHEUTZ, F., LUND, O., HASMAN, H., KAAS, R. S., NIELSEN, E. M. & AARESTRUPA, F. M. 2014b. Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli. *Journal of Clinical Microbiology* 52**,** 1501–1510s.

JOLLEY, K. A., BRAY, J. E. & MAIDEN, M. C. J. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res,* 3**,** 124.

JUAN, C., PENA, C. & OLIVER, A. 2017. Host and Pathogen Biomarkers for Severe Pseudomonas aeruginosa Infections. *J Infect Dis,* 215**,** S44-S51.

KAN, B., ZHOU, H., DU, P., ZHANG, W., LU, X., QIN, T. & XU, J. 2018b. Transforming bacterial disease surveillance and investigation using whole-genome sequence to probe the trace. *Front Med,* 12**,** 23-33.

KERR, K. G. & SNELLING, A. M. 2009. Pseudomonas aeruginosa: a formidable and ever-present adversary. *J Hosp Infect,* 73**,** 338-44.

KLOCKGETHER, J. & TUMMLER, B. 2017. Recent advances in understanding Pseudomonas aeruginosa as a pathogen. *F1000Res,* 6**,** 1261.

KOVANEN, S. M., KIVISTÖ, R. I., ROSSI, M., SCHOTT, T., KÄRKKÄINEN, U.-M., TUUMINEN, T., UKSILA, J., RAUTELIN, H. & HÄNNINENA, M.-L. 2014b. Multilocus Sequence Typing (MLST) and Whole-Genome MLST of Campylobacter jejuni Isolates from Human Infections in Three Districts during a Seasonal Peak in Finland. *Journal of Clinical Microbiology* 52**,** 4147–4154.

KRAMER, A., SCHWEBKE, I. & KAMPF, G. 2006. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC Infect Dis,* 6**,** 130.

LIVERMORE, D. M. 2002. Multiple Mechanisms of Antimicrobial Resistance in Pseudomonas aeruginosa: Our Worst Nightmare? *Clinical Infectious Diseases,* 34**,** 634–40.

LODISE, T. P., JR., PATEL, N., KWA, A., GRAVES, J., FURUNO, J. P., GRAFFUNDER, E., LOMAESTRO, B. & MCGREGOR, J. C. 2007. Predictors of 30-day mortality among patients with Pseudomonas aeruginosa bloodstream infections: impact of delayed appropriate antibiotic selection. *Antimicrob Agents Chemother,* 51**,** 3510-5.

MAIDEN, M. C. J., FEIL, J. A. B. E., MORELL, G., URWIN, J. E. R. R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M., SPRATT, A. B. G., BYGRAVES, J. A., FEIL, E., MORELL, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M. & SPRATT, B. G. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* 95**,** 3140–3145.

MCNALLY, A., OREN, Y., KELLY, D., PASCOE, B., DUNN, S., SREECHARAN, T., VEHKALA, M., VALIMAKI, N., PRENTICE, M. B., ASHOUR, A., AVRAM, O., PUPKO, T., DOBRINDT, U., LITERAK, I., GUENTHER, S., SCHAUFLER, K., WIELER, L. H., ZHIYONG, Z., SHEPPARD, S. K., MCINERNEY, J. O. & CORANDER, J. 2016. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet,* 12**,** e1006280.

MIHARA, T., KIMURA, T. & AL., K. M. E. 2020. Secondary in-hospital epidemiological investigation after an outbreak of Pseudomonas aeruginosa ST357. *Journal of Infection and Chemotherapy*, 26.

MOORE, L. S., FREEMAN, R., GILCHRIST, M. J., GHARBI, M., BRANNIGAN, E. T., DONALDSON, H., LIVERMORE, D. M. & HOLMES, A. H. 2014. Homogeneity of antimicrobial policy, yet heterogeneity of antimicrobial resistance: antimicrobial non-susceptibility among 108,717 clinical isolates from primary, secondary and tertiary care patients in London. *J Antimicrob Chemother,* 69**,** 3409-22.

MOURA, A., CRISCUOLO, A., POUSEELE, H., MAURY, M. M., LECLERCQ, A., TARR, C., BJORKMAN, J. T., DALLMAN, T., REIMER, A., ENOUF, V., LARSONNEUR, E., CARLETON, H., BRACQ-DIEYE, H., KATZ, L. S., JONES, L., TOUCHON, M., TOURDJMAN, M., WALKER, M., STROIKA, S., CANTINELLI, T., CHENAL-FRANCISQUE, V., KUCEROVA, Z., ROCHA, E. P., NADON, C., GRANT, K., NIELSEN, E. M., POT, B., GERNER-SMIDT, P., LECUIT, M. & BRISSE, S. 2016. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nat Microbiol,* 2**,** 16185.

PARCELL, B. J., ORAVCOVA, K., PINHEIRO, M., HOLDEN, M. T. G., PHILLIPS, G., TURTON, J. F. & GILLESPIE, S. H. 2018. Pseudomonas aeruginosa intensive care unit outbreak: winnowing of transmissions with molecular and genomic typing. *J Hosp Infect,* 98**,** 282-288.

PEARCE, M. E., ALIKHAN, N. F., DALLMAN, T. J., ZHOU, Z., GRANT, K. & MAIDEN, M. C. J. 2018. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. *Int J Food Microbiol,* 274**,** 1-11.

PELEGRIN, A. C., GRIFFON, Y. R. S. A., PALMIERI, M., MIRANDE, C., KARUNIAWATI, A., SEDONO, R., ADITIANINGSIH, D., GOESSENS, W. H. F., BELKUM, A. V., VERBRUGH, H. A., KLAASSEN, C. H. W. & SEVERIND, J. A. 2019. High-risk international clones of carbapenem-nonsusceptible Pseudomonas aeruginosa endemic to Indonesian intensive care units: impact of a multifaceted infection control intervention analyzed at the genomic level. *mBio,* 10.

PELEGRIN AC, S. Y., GRIFFON A, PALMIERI M, MIRANDE C, KARUNIAWATI A, SEDONO R, ADITIANINGSIH D, GOESSENS WHF, VAN BELKUM A, VERBRUGH HA, KLAASSEN CHW, SEVERIN JA. 2019. High-risk international clones of carbapenem-nonsusceptible Pseudomonas aeruginosa endemic to Indonesian intensive care units: impact of a multifaceted infection control intervention analyzed at the genomic level. *mBio* 10**,** e02384-19.

PUJA, H., BOLARD, A., NOGUÈS, A., PLÉSIAT, P. & JEANNOT, K. 2020. The efflux pump MexXY/OprM contributes to the tolerance and acquired resistance of Pseudomonas aeruginosa to colistin. *Antimicrob Agents Chemother* 64**,** e02033-19.

QUICK, J., CUMLEY, N., WEARN, C. M., NIEBEL, M., CONSTANTINIDOU, C., THOMAS, C. M., PALLEN, M. J., MOIEMEN, N. S., BAMFORD, A., OPPENHEIM, B. & LOMAN, N. J. 2014. Seeking the source of Pseudomonas aeruginosa infections in a recently opened hospital: an observational study using whole-genome sequencing. *BMJ Open,* 4**,** e006278.

ROE, C. C., HORN, K. S., DRIEBE, E. M., BOWERS, J., TERRIQUEZ, J. A., KEIM, P. & ENGELTHALER, D. M. 2016. Whole genome SNP typing to investigate methicillin-resistant Staphylococcus aureus carriage in a health-care provider as the source of multiple surgical site infections. *Hereditas,* 153**,** 11.

RUMORE, J., TSCHETTER, L., KEARNEY, A., KANDAR, R., MCCORMICK, R., WALKER, M., PETERSON, C. L., REIMER, A. & NADON, C. 2018b. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic Escherichia coli O157:H7 from the Canadian perspective. *BMC Genomics,* 19**,** 870.

SAHARMAN, Y., PELEGRIN, A., KARUNIAWATI, A., SEDONO, R., ADITIANINGSIH, D., GOESSENS, W., KLAASSEN, C., BELKUM, A. V., MIRANDE, C., VERBRUGH, H. & SEVERIN, J. 2019 Epidemiology and characterisation of carbapenem-non-susceptible Pseudomonas aeruginosa in a large intensive care unit in Jakarta, Indonesia. *Int J Antimicrob Agents. ,* 54**,** 655-660. .

SCHURCH, A. C., ARREDONDO-ALONSO, S., WILLEMS, R. J. L. & GOERING, R. V. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clin Microbiol Infect,* 24**,** 350-354.

STANTON, R. A., MCALLISTER, G., DANIELS, J. B., BREAKER, E., VLACHOS, N., GABLE, P., MOULTON-MEISSNER, H. & HALPIN, A. L. 2020. Development and Application of a Core Genome Multilocus Sequence Typing Scheme for the Health Care-Associated Pathogen Pseudomonas aeruginosa. *J Clin Microbiol,* 58.

STOVER, C. K., PHAM, X. Q., ERWIN, A. L., MIZOGUCHI, S. D., WARRENER, P., HICKEY, M. J., BRINKMAN, F. S. L., HUFNAGLE, W. O., KOWALIK, D. J., LAGROU, M., GARBER, R. L., GOLTRY, L., TOLENTINO, E., WESTBROCK-WADMAN, S., YUAN, Y., BRODY, L. L., COULTER, S. N., FOLGER, K. R., KAS, A., LARBIG, K., R.LIM, SMITH, K., SPENCER, D., WONG, G. K.-S., Z.WU, PAULSENK, I. T., REIZER, J., SAIER, M. H., LORY, R. E. W. H. S. & OLSON, M. V. 2000. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *NATURE* 406

STUDEMEISTER, A. E. & QUINN, J. P. 1988. Selective Imipenem Resistance in Pseudomonas aeruginosa Associated with Diminished Outer Membrane Permeability. *ANTIMICROBIAL AGENTS AND CHEMOTHERAPY,* 32**,** 1267-1268.

SUBEDI, D., KOHLI, G. S., VIJAY, A. K., WILLCOX, M. & RICE, S. A. 2019. Accessory genome of the multi-drug resistant ocular isolate of Pseudomonas aeruginosa PA34. *PLoS One,* 14**,** e0215038.

SURESH, M., SKARIYACHAN, S., NARAYANAN, N., PULLAMPARA RAJAMMA, J. & PANICKASSERY RAMAKRISHNAN, M. K. 2020. Mutational Variation Analysis of oprD Porin Gene in Multidrug-Resistant Clinical Isolates of Pseudomonas aeruginosa. *Microb Drug Resist,* 26**,** 869-879.

TADA, T., T, H., S, W., H, U., M, T., K, K.-A., S, M., KN, Z., T, K. & HH, T. 2019. Molecular Characterization of Multidrug-Resistant Pseudomonas aeruginosa Isolates in Hospitals in Myanmar. *Antimicrobial Agents and Chemotherapy,* 63 e02397-18.

TAYLOR, A. J., LAPPI, V., WOLFGANG, W. J., LAPIERRE, P., PALUMBO, M. J., MEDUS, C. & BOXRUD, D. 2015. Characterization of Foodborne Outbreaks of Salmonella enterica Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection. *J Clin Microbiol,* 53**,** 3334-40.

TREEPONG, P., V.N.KOS, C.GUYEUX, D.S.BLANC, X.BERTRAND, B.VALOT & D.HOCQUET 2018. Global emergence of the widespread Pseudomonas aeruginosa ST235 clone. *Clinical Microbiology and Infection*
24**,** 258-266.

URWIN, R. & MAIDEN, M. C. J. 2003. Multi-locus sequence typing:a tool for global epidemiology. *TRENDS in Microbiology*, 11.

YIN, C. & YAU, S. S.-T. 2018. Whole genome single nucleotide polymorphism genotyping of Staphylococcus aureus. arXiv.

YONG, W., GUO, B., SHI, X., CHENG, T., CHEN, M., JIANG, X., YE, Y., WANG, J., XIE, G. & DING, J. 2018. An Investigation of an Acute Gastroenteritis Outbreak: Cronobacter sakazakii, a Potential Cause of Food-Borne Illness. *Front Microbiol,* 9**,** 2549.

ZANKARI, E., HASMAN, H., COSENTINO, S., VESTERGAARD, M., RASMUSSEN, S., LUND, O., AARESTRUP, F. M. & LARSEN, M. V. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother,* 67**,** 2640-4.

ZOWAWI, H. M., SYRMIS, M. W., KIDD, T. J., BALKHY, H. H., WALSH, T. R., AL JOHANI, S. M., AL JINDAN, R. Y., ALFARESI, M., IBRAHIM, E., AL-JARDANI, A., AL SALMAN, J., DASHTI, A. A., SIDJABAT, H. E., BAZ, O., TREMBIZKI, E., WHILEY, D. M. & PATERSON, D. L. 2018. Identification of carbapenem-resistant Pseudomonas aeruginosa in selected hospitals of the Gulf Cooperation Council States: dominance of high-risk clones in the region. *J Med Microbiol,* 67**,** 846-853.

# Chapter 6

## Epidemiological analysis of *Pseudomonas aeruginosa* using EPISEQ® CS: an advanced one-stop solution for Next Generation Sequencing data analysis

Manisha Goyal[1], Benjamin Moingeon[2], Stephane Bulteau[3], Jill Dombrecht[4], Katrien De Bruyne[4] and Alex van Belkum[1]*

[1]bioMérieux Open innovation and Partnerships, 3 Route du Port Michaud, 038391 La Balme Les Grottes, France

[2]bioMérieux EU Data Science, 376 Chemin de l'Orme, 69280, Marcy L'Etoile, France

[3]bioMérieux Lab Informatics & Analytics, Grenoble, France

[4]bioMérieux Applied Maths NV, 9830 Sint-Martens-Latem, Belgium

**Abstract**

Next Generation Sequencing (NGS) is rapidly becoming the Gold Standard method for the epidemiological tracing of bacterial pathogens. Whole Genome Sequences (WGS) provide a wealth of information on genomic identity of bacterial strains that were isolated from the same clinical environment. This information can be used to define whether or not strains share a common origin. This can have a major impact on nosocomial infection control but in order to do so, the WGS data interpretation should be simplified and made accessible to non-bio-informaticians in an easy and straightforward manner. We here present EPISEQ® CS as a tool for rapid translation of primary WGS data into actionable advice for hospital-based microbiologists and infection control professionals.

Using WGS for *Pseudomonas aeruginosa* as an example, here we carried out preassembly quality assessment of reads, *de novo* genome assembly, comparative strain characterization at the WGS level, AMR gene profiling and phylogenetic analysis at a push-button level using EPISEQ® CS. Also compared that with other pipelines like bioNumerics. Similar results with a few advantages as well as disadvantages of the two different pipelines were observed. Unlike other available WGS data analysis pipelines EPISEQ® CS works as an automated system for epidemiological genome analysis, does not require bio-informatic expertise and provides a full consolidated output report. However some parametric access would be a plus in order to improve the quality and efficiency of EPISEQ® CS.

## Introduction

Health care-associated bacterial infections are one of the leading causes of nosocomial morbidity and mortality worldwide (Genovese et al., 2020). Each year between 400,000 and 720,000 cases of healthcare-associated infections (HAI) are estimated to occur in the US only (Magill et al., 2014, Zimlichman et al., 2013). *Pseudomonas aeruginosa*, as one of the major pathogens of HAI, is associated with substantially higher mortality and morbidity rates than those defined for other pathogens (Dellinger, 2016). Through mutation and acquisition of resistance elements, *P. aeruginosa* has developed populations that are well adapted to the local use of antiseptics and antibiotics. According to the antimicrobial resistance (AMR) threat report published in 2019, multidrug-resistant (MDR) *P. aeruginosa* caused an estimated 32,600 infections among hospitalized patients and 2,700 estimated deaths in the US in the year 2017 (Health and Services, 2019)

Microbiological detection methods, more precise strain characterization strategies and epidemiological analyses have evolved significantly beyond the classical and mostly phenotypic methodologies (Zeeshan and Razzak, 2020). The first step in the epidemiological analysis of microbial isolates comprises experimental typing of microorganisms, which was historically done using a broad variety of conventional methods. The most widely used classical strain characterization techniques were PCR-based using among others the amplification and sizing of variable numbers of tandem repeats (VNTRs), restriction fragment length polymorphisms (RFLP) and amplified fragment length polymorphisms (AFLP) techniques (Manukumar and Umesha, 2017). More recently, electrophoretic and PCR-based typing became the global methods of choice (Shokoohizadeh, 2016, Sánchez, 2015). Later, mass spectroscopy (MS) was developed for bacterial identification mostly, but applications in the field of typing were reported as well. These technologies are different from each other in terms of discriminatory power, reproducibility, timelines, portability and cost effectiveness (Babalola, 2003, Lasker, 2002, Van Belkum et al., 2007). Initially and despite its high resolution, the value of whole genome sequencing (WGS) was underestimated due to cumbersome methodology and high costs (Margulies et al., 2005, Valouev et al., 2008, Quainoo et al., 2017, Rothberg et al., 2011). Still, next generation sequencing (NGS) that facilitates WGS is now considered the new Gold Standard typing methodology.

Recent technological advancements in NGS technology further pushed the use of WGS in the field of infectious disease research and especially for the analyses of infectious outbreaks and pathogen surveillance (Van Goethem et al., 2019, Quainoo et al., 2017, Goldberg et al., 2015, Dylus et al., 2020). Several bio-informatic tools and algorithms can now be utilized for NGS data analysis with as main foci overall genomic strain characterization, detection of existing and new antimicrobial resistance (AMR) genes and virulence factors, and definition of strain transmission dynamics in health care settings (Van Goethem et al., 2019, Tshibangu-Kabamba et al., 2020, Goyal et al., 2020, Goldberg et al., 2015, Dylus et al., 2020) In this highly competitive scientific field many different data analysis software packages (e.g. sraX (https://github.com/lgpdevtools/srax; (Panunzi, 2020)), BacPipe (Xavier et al., 2020), CLC Genomics Workbench by QIAGEN (https://digitalinsights.qiagen.com), BIONUMERICS (Applied Maths, bioMérieux)) are commercially or freely available. The software service called EPISEQ® CS (bioMérieux, Marcy L'Etoile, France) is one of the recently developed NGS data analysis tools. The system is based on interactive graphical user interfaces (GUI) and aims at non-specialist users. It generates an integral report providing a complete epidemiological analysis along with a graphical phylogenetic tree, a minimum spanning tree and a quality check for the raw data and the resulting genomic assembly. Most importantly, it generates a full epidemiological analysis report with automatic color coding for various themes and metadata. Individual sample reports can also be accessed, indicating the complete resistome and virulome for an individual strain.

The present study focuses on the epidemiological analysis of a panel of 214 *P. aeruginosa* strains collected from a single Indonesian hospital during an infection control intervention (Pelegrin AC, 2019). Here, we have specifically exploited raw sequence reads and our main focus was to define the efficiency and reliability of EPISEQ® CS in comparison with alternative data interpretation pipelines.

## Material and Methods

Raw sequence reads for the set of 214 *P. aeruginosa* strains from a major Indonesian hospital in Jakarta were used for epidemiological analysis before (Pelegrin AC, 2019). The Illumina FASTQ reads were now uploaded to a personal account in the bioMérieux EPISEQ® CS software application available at https://data-analytics.biomerieux.com. After

trimming of raw reads, de novo assembly was performed using SPAdes v3.10.0. During post processing, a consensus assembly was generated and quality metrics on raw and trimmed data reads and the assemblies were calculated. These quality metrics were compared with the pre-defined thresholds in EPISEQ® CS. Unlike done in the present study, assembled FASTA files instead of raw reads with a minimal recommended coverage of 45x and a minimal paired end read length of 150 base pairs can also be directly uploaded in EPISEQ® CS. Quality parameters such as the number of core loci present, custom k-mer classification of assembled genomes and assembled genome lengths were checked in order to verify the authenticity of the uploaded genome sequence data. Dedicated modules in EPISEQ® CS determined the contamination score based on the presence of non-ATGC bases aided by the k-mer classification of assembled genomes. Thereafter, epidemiological analysis was done using a whole genome multi-locus sequence typing (wgMLST) approach. CARD (version 1.2.1; 2017-10-10; (Jia et al., 2016)), ResFinder (2019-09-16; (Zankari et al., 2012)), PointFinder (2019-09-16; https://bitbucket.org/genomicepidemiology/pointfinder_db/) and NCBI antimicrobial resistance databases (version: 2020-06-11; https://www.ncbi.nlm.nih.gov/pathogens/isolates#/refgene/) were aggregated and reformatted to fit the EPISEQ® CS resistance ontology pathway. EPISEQ® CS generates a set of allele differences that describe the degree of dissimilarity between individual strains that can be represented as a minimum spanning tree (MST) which ultimately defines the most likely chain of pathogen transmission. The MST developed in EPISEQ® CS is build using the Prim algorithm (https://gowalker.org/github.com/soniakeys/graph#LabeledUndirected_Prim). The phylogenetic dendrogram is based upon the unweighted pair group method of arithmetic mean (UPGMA) method of cluster analysis.

**Results and Discussion**

Being an automated system for epidemiological genome analysis, EPISEQ® CS does not require bio-informatic expertise. Still, it needs a molecularly trained infection control professional to execute and interpret the full workflow. Unlike otherwise available WGS data analysis pipelines, EPISEQ® CS provides a full report including preassembly quality assessment of reads, *de novo* genome assembly, comparative strain characterization at the WGS level, AMR gene profiling and phylogenetic analysis at a push-button level. Apart from the overall epidemiologically oriented report, EPISEQ® CS generates individual reports for

all strains. Consequently, the report describes the strain-specific resistome and hints at antimicrobial drugs to which strains are expected to be susceptible or resistant.

Population-based Relationship Analysis

One of the factors that sets EPISEQ® CS apart from other pipelines is the complete population-based analysis it performs using data from all uploaded strains. Relationships based on wgMLST similarities within any new dataset as well as between the new dataset and the WGS data already available in the overall database of EPISEQ® CS are automatically calculated (Figure 6-1). Within the input dataset of the 214 Indonesian *P. aeruginosa* genome sequences studied in this communication, 22 strains with <= 94.91% similarity (10.3%) were identified as unrelated and unique, 52 strains sharing similarities between 95% to 98.99% (24.3%) were considered as possibly related and 140 strains with >= 99.89% similarity (65.4.7%) were contemplated as probably related. However, when the Indonesian *P. aeruginosa* WGS where compared to the entire EPISEQ® CS database only 43 genome sequences were found to be probably related with entries in the global database. This points to the circulation of a relatively unique set of strains in the ICU setting of the Indonesian hospital. Consequently, only a small number of entries in the database display elevated levels of homology with the genomes of the Indonesian strains.

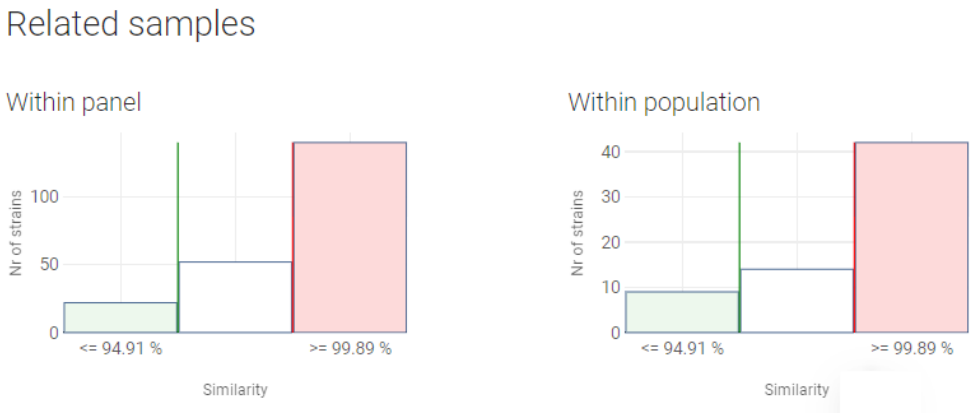



***Figure 6-1:*** *Relationship analysis of the genomes of indicidual Indonesian P. aeruginosa strains with all the other strains in the Indonesian input panel (within panel, box on the left) as well as with those avaiable in EPISEQ® CS database (within population, box on the right). The green colored bar represents the number of strains found unrelated, the white bar in the middle of both of the graphs shows the number of possibly related samples and red colored bar determines the number of probably related strains. Similarity figures are shown on horizontal axis.*

Quality Control

Initially, identity and read quality were checked for each sample. Then at execution of each step of the pipeline, additional quality analysis was performed. Paired end raw as well as trimmed reads were examined on the basis of prefixed parameters which defines the probability of incorrect base calls at the termini of raw reads. This section of the report explains the overall quality of the data by giving color indications of major and minor warnings (see Figure 6-2 for an example). The QC report generated during the present analysis demonstrates that the quality of pre-intervention *P. aeruginosa* strains was slightly better than that of post-intervention *P. aeruginosa* strains (Supplementary file).

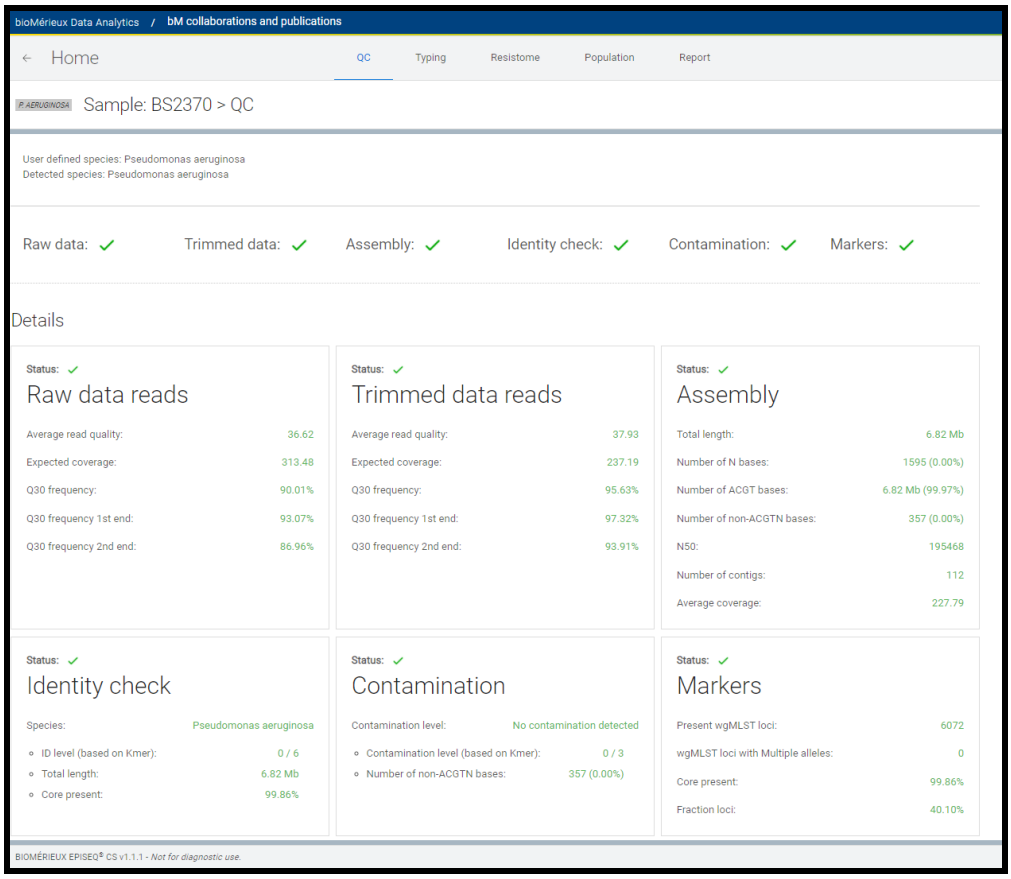



***Figure 6-2:*** *Quality control measures (sample BS2370) carried out at each step of the analysis strating from raw data reads to markers calling. Status of good quality data is color coded as green where as yellow and red status are the sign of minor and major warnings respectively appears during any step of the analysis.*

wgMLST-based Strain Typing and Phylogenetic Analyses

*De novo* assembled genomes were characterized using thousands of loci scattered around the *P. aeruginosa* genomes. Out of 214 genomes, 198 MLST sequence types were accurately identified by seven locus MLST, in comparison with a similar analysis done by (Pelegrin AC, 2019), using a prior version of the MLST database, adequate MLST concordance was observed (Supplementary file). Next to MLST, phylogenetic analysis based on the wgMLST profiles provides more detailed insight into the level of strain relatedness. For each species there are pre-determined similarity thresholds for calling probable or possible relatedness. A historical database for each species has been built over time and can be used to compare bacterial strains from new or current outbreaks with those from previous events. A commonly used tool for short term molecular epidemiology of bacterial strains involves the use of MSTs. Between the strains of the Indonesian input panel, allele differences ranged from 0 to 3400 and using a distance matrix, an MST was constructed. Developing the MST requires mathematically sophisticated algorithms such as phylogenetic analysis for inferring genetically diverse population structures. Here we generated MSTs using the wgMLST data of the same set of strains of two major groups ST235 and ST357 with BIONUMERICS as well as with EPISEQ® CS. The output (Figure 6-3) illustrates that allelic variations (SLVs, DLVs and TLVs) among the strains were more clearly defined in the EPISEQ® CS generated tree. In contrast, only wgMLST-based groupings were formed in the BIONUMERICS generated tree.

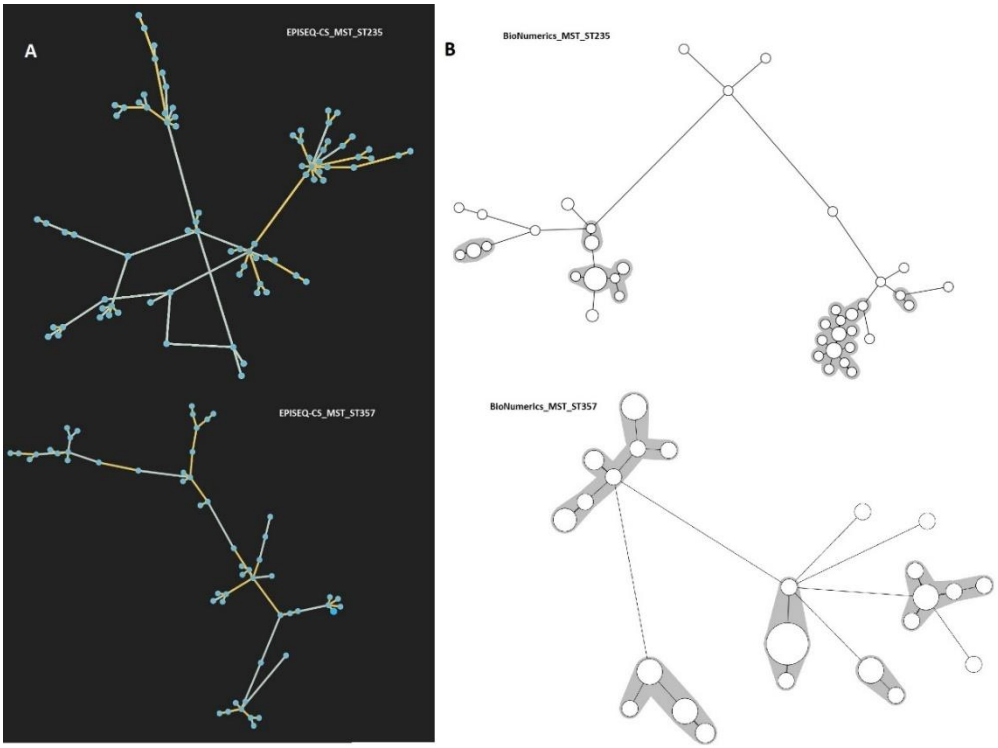



***Figure 6-3:*** *Minimum spanning tree calculated from an input panel of P. aeruginosa strains from Indonesia. Image A represents MST based on allelic variations in ST235 and ST357 generated by EPISEQ® CS (in black background) and Image B represents the same generated by BIONUMERICS (In white background). Allelic variations are not shown in image A.*

The epidemiological analysis window of EPISEQ® CS provides a dendrogram along with metadata and quality control parameters of the input panel. Color-coded clustering on the basis of wgMLST can also be seen in Figure 6-4. In the present analysis, EPISEQ® CS generated a phylogenetic tree that demonstrated the clustering of input panel strains including the outside population database developed by EPISEQ® CS from previously uploaded strains (Figure 6-4). Using predefined UPGMA, thirty two clusters at 99.89 % similarity (by default) were identified in this study. At any time, one can identify the number of clusters based on custom defined similarity thresholds, providing a highly interactive manner for the user to characterize the relationship between strains according to study requirements.

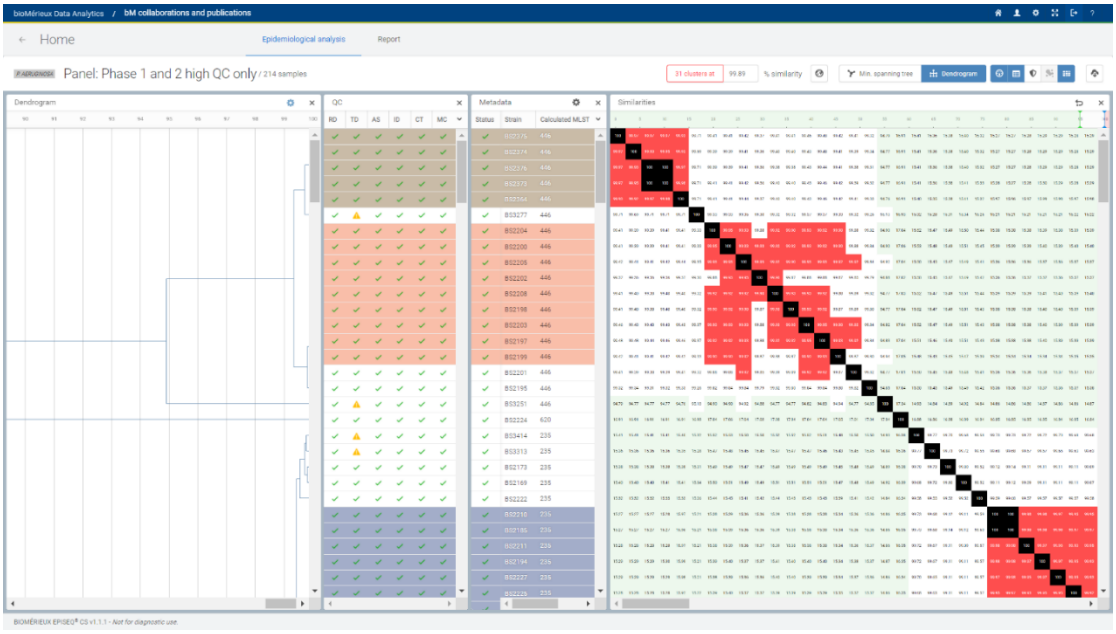



***Figure 6-4:*** *Epidemiological analysis window of EPISEQ® CS showing the dendrogram, QC, metadata and the similarity matrix generated for P. aeruginosa strains from Indonesia using the EPISEQ® CS database.*

EPISEQ® CS does not allow to change the parameters to generate dendrograms in the way BIONUMERICS does. Considering this limitation we have here generated a wgMLST based dendrogram using both BIONUMERICS and EPISEQ® CS (Figure 6-5). Clustering was almost identical, supporting the efficiency of the EPISEQ® CS output. Other strains from the EPISEQ® CS database sharing more than 99.89% similarity with the Indonesian input panel were also analysed. In the present analysis our input panel formed 50 clusters with strains from the EPISEQ® CS knowledge base. Individual dendrograms were constructed for each strain to explain its relationship with database strains. This could potentially map the historical origin of a strain or to find its descendants in previous outbreaks.

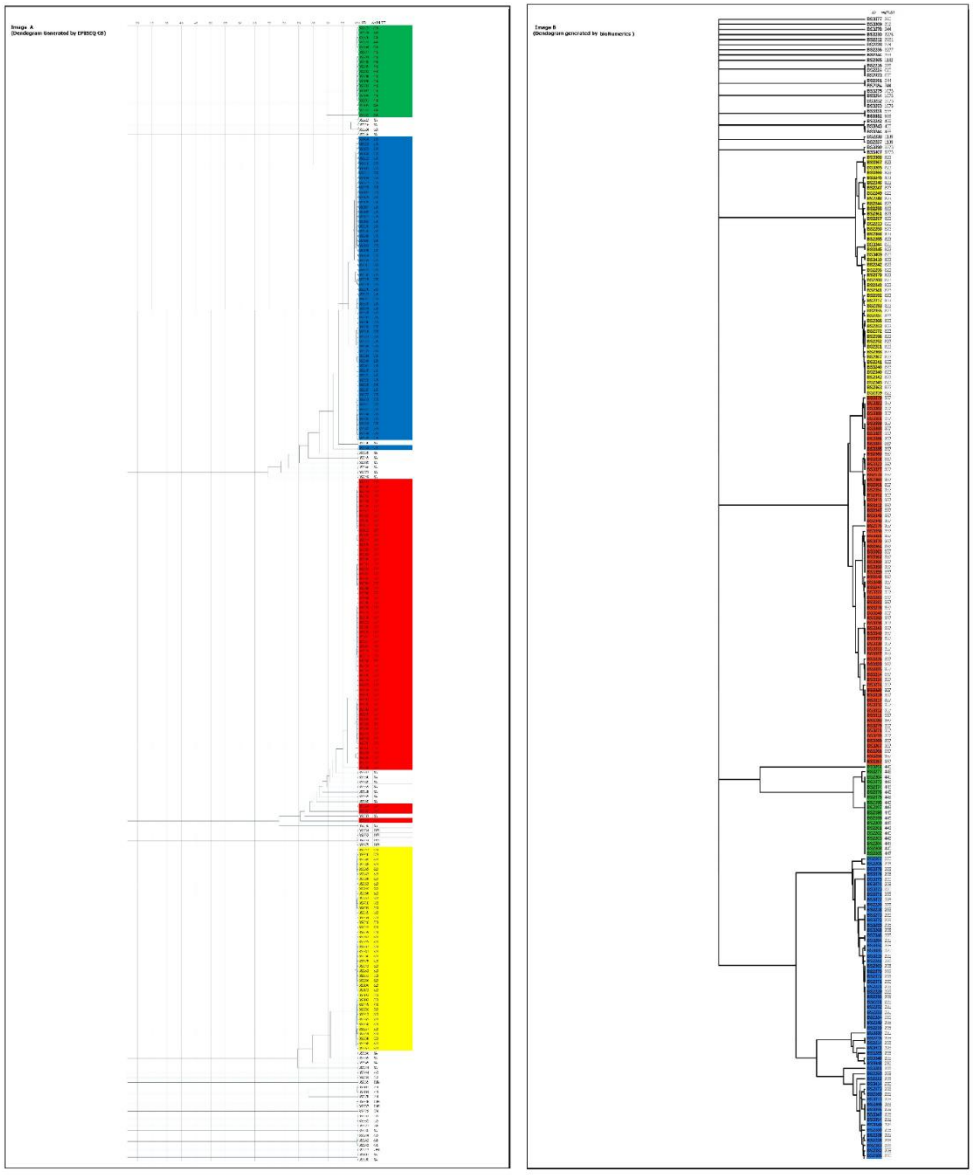


***Figure 6-5:*** *Image A: dendrogram generated by EPISEQ® CS. Image B: dendrogram generated with BIONUMERICS along with strain ID and wgMLST.*

Detailed Resistome Analysis

Unlike most epidemiology pipelines, resistome analysis in EPISEQ® CS is not limited to mere detection of resistance genes; It also provides a detailed exploration of resistance alleles including the identity, coverage, position of the allele in the genome, information of drugs to which a resistance gene may confer resistance, drug families and the resistance mechanism provided by the resistance marker (see Figure 6-6 for an example). We found SNPs within the resistome of each WGS, both known ones and ones newly discovered. This further supported the information on therapeutic usefulness of antimicrobial drugs with

which the respected AMR marker associates (Figure 6- 6). Other resistance finder tools or pipelines provide similar information on resistance genes in the genome along with the identity percentages, ARG associated mutations or SNPs. Still, the recently developed tool sraX for instance is not a one-step application such as EPISEQ® CS, but rather a complex command line tool (Panunzi, 2020). For comparative purposes, we performed the resistome analysis using the same dataset and another command line-based software called Abricate (https://github.com/tseemann/ABRicate; (Sydenham et al., 2019)) generating a list with ARGs and their identity coverage, homology percentage, comprehensive list of mutations and resistance to certain drugs and the mechanism of resistance action. EPISEQ® CS (Version 2.0), which provides tabular information on resistome content and mutations, was simple to perform and its outcome was easy to interpret and understand (Supplementary file). However, no virulome information was obtained for the input panel. Other dedicated tools for epidemiology i.e. bacPipe and sarX are highly flexible in terms of user defined parameters but none of those provides such elaborate information on antimicrobial resistance based on specific mutations as EPISEQ® CS does.

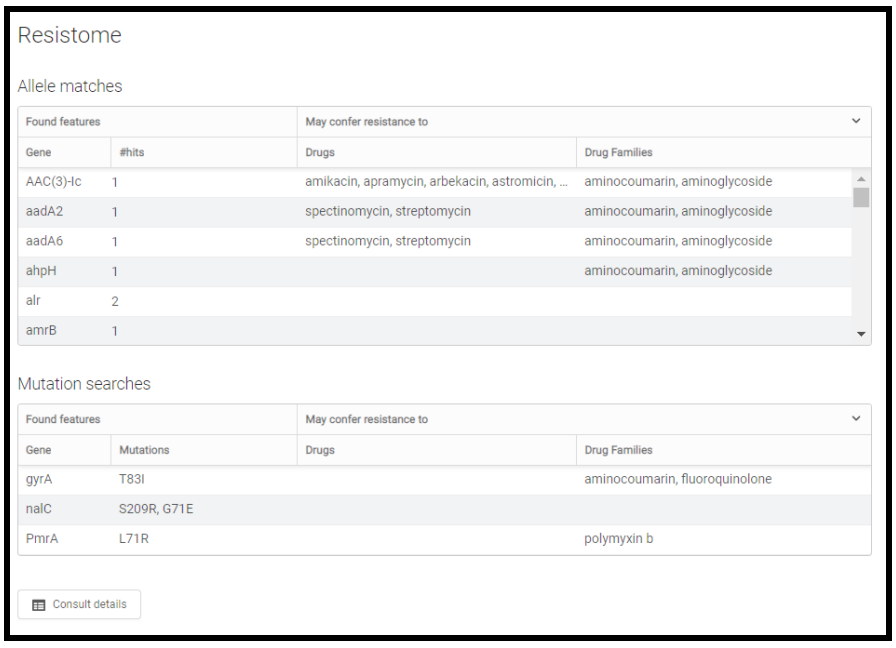



***Figure 6-6:*** *Resistome analysis for strain BS2370 in EPISEQ® CS with all mutations identified in its resistome along with their mechanism of action.*

## Conclusion

EPISEQ® CS is an advanced bio-informatic pipeline primarily focused on epidemiological strain characterization but also including the identification of resistome markers. It is a push-button solution as compared to other available NGS data analysis tools. It does not require extensive expertise in bio-informatics, command line platforms and technical knowledge of background algorithms to define the epidemiological and resistome-associated parameters. Although it takes time and high bandwidth internet connection to upload the hundreds of sequence datasets, it does not take much time to complete the full analysis using the largely predefined parameters. Moreover, results are well organized in the form of color coded figures and tables. Identification of new mutations in a resistome filters for novel markers. EPISEQ® CS is an interesting tool in a field that is currently dominated by academic researchers and diagnostic laboratories in first tier hospitals. Simple and push-button tools such as EPISEQ® CS will provide wider access to this technology.

## Supplementary file

MLST of the dataset of *Pseudomonas aeruginosa* strains calculated by EPISEQ® CS.

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

Samples (# 235)

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 1 | BS3414 | Pseudomonas aeruginosa | 235 |
| 2 | BS3413 | Pseudomonas aeruginosa | 357 |
| 3 | BS3412 | Pseudomonas aeruginosa | 357 |
| 4 | BS3411 | Pseudomonas aeruginosa | 235 |
| 5 | BS3410 | Pseudomonas aeruginosa | 823 |
| 6 | BS3409 | Pseudomonas aeruginosa | 823 |
| 7 | BS3407 | Pseudomonas aeruginosa | NA |
| 8 | BS3390 | Pseudomonas aeruginosa | NA |
| 9 | BS3389 | Pseudomonas aeruginosa | 357 |
| 10 | BS3388 | Pseudomonas aeruginosa | 357 |
| 11 | BS3387 | Pseudomonas aeruginosa | 357 |
| 12 | BS3386 | Pseudomonas aeruginosa | 357 |
| 13 | BS3385 | Pseudomonas aeruginosa | 357 |
| 14 | BS3384 | Pseudomonas aeruginosa | 357 |
| 15 | BS3383 | Pseudomonas aeruginosa | 357 |
| 16 | BS3382 | Pseudomonas aeruginosa | 357 |
| 17 | BS3381 | Pseudomonas aeruginosa | 357 |



Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 18 | BS3380 | Pseudomonas aeruginosa | 357 |
| 19 | BS3379 | Pseudomonas aeruginosa | 357 |
| 20 | BS3378 | Pseudomonas aeruginosa | 235 |
| 21 | BS3377 | Pseudomonas aeruginosa | 260 |
| 22 | BS3376 | Pseudomonas aeruginosa | 235 |
| 23 | BS3375 | Pseudomonas aeruginosa | 235 |
| 24 | BS3374 | Pseudomonas aeruginosa | 235 |
| 25 | BS3373 | Pseudomonas aeruginosa | NA |
| 26 | BS3372 | Pseudomonas aeruginosa | 235 |
| 27 | BS3371 | Pseudomonas aeruginosa | 235 |
| 28 | BS3370 | Pseudomonas aeruginosa | 357 |
| 29 | BS3369 | Pseudomonas aeruginosa | 312 |
| 30 | BS3368 | Pseudomonas aeruginosa | 823 |
| 31 | BS3367 | Pseudomonas aeruginosa | 823 |
| 32 | BS3366 | Pseudomonas aeruginosa | NA |
| 33 | BS3365 | Pseudomonas aeruginosa | NA |
| 34 | BS3364 | Pseudomonas aeruginosa | 357 |

For research use only. Not for use in diagnostic procedures.
Copyright © 2020 - 2021 BIOMÉRIEUX EPISEQ® CS v1.1.0 - All Rights Reserved

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 35 | BS3363 | Pseudomonas aeruginosa | NA |
| 36 | BS3362 | Pseudomonas aeruginosa | 357 |
| 37 | BS3361 | Pseudomonas aeruginosa | 446 |
| 38 | BS3360 | Pseudomonas aeruginosa | NA |
| 39 | BS3359 | Pseudomonas aeruginosa | NA |
| 40 | BS3358 | Pseudomonas aeruginosa | 357 |
| 41 | BS3356 | Pseudomonas aeruginosa | NA |
| 42 | BS3355 | Pseudomonas aeruginosa | NA |
| 43 | BS3354 | Pseudomonas aeruginosa | 235 |
| 44 | BS3350 | Pseudomonas aeruginosa | NA |
| 45 | BS3349 | Pseudomonas aeruginosa | NA |
| 46 | BS3346 | Pseudomonas aeruginosa | NA |
| 47 | BS3345 | Pseudomonas aeruginosa | NA |
| 48 | BS3344 | Pseudomonas aeruginosa | NA |
| 49 | BS3341 | Pseudomonas aeruginosa | 357 |
| 50 | BS3340 | Pseudomonas aeruginosa | NA |
| 51 | BS3339 | Pseudomonas aeruginosa | NA |



Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 52 | BS3338 | Pseudomonas aeruginosa | NA |
| 53 | BS3337 | Pseudomonas aeruginosa | NA |
| 54 | BS3336 | Pseudomonas aeruginosa | NA |
| 55 | BS3335 | Pseudomonas aeruginosa | NA |
| 56 | BS3334 | Pseudomonas aeruginosa | 235 |
| 57 | BS3333 | Pseudomonas aeruginosa | 357 |
| 58 | BS3332 | Pseudomonas aeruginosa | 555 |
| 59 | BS3331 | Pseudomonas aeruginosa | 555 |
| 60 | BS3330 | Pseudomonas aeruginosa | 235 |
| 61 | BS3329 | Pseudomonas aeruginosa | 357 |
| 62 | BS3328 | Pseudomonas aeruginosa | 357 |
| 63 | BS3327 | Pseudomonas aeruginosa | 357 |
| 64 | BS3326 | Pseudomonas aeruginosa | 357 |
| 65 | BS3325 | Pseudomonas aeruginosa | 357 |
| 66 | BS3324 | Pseudomonas aeruginosa | 357 |
| 67 | BS3323 | Pseudomonas aeruginosa | 357 |
| 68 | BS3322 | Pseudomonas aeruginosa | 357 |



Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 69 | BS3321 | Pseudomonas aeruginosa | 357 |
| 70 | BS3320 | Pseudomonas aeruginosa | 357 |
| 71 | BS3319 | Pseudomonas aeruginosa | 357 |
| 72 | BS3317 | Pseudomonas aeruginosa | 357 |
| 73 | BS3316 | Pseudomonas aeruginosa | 357 |
| 74 | BS3315 | Pseudomonas aeruginosa | 235 |
| 75 | BS3314 | Pseudomonas aeruginosa | 357 |
| 76 | BS3313 | Pseudomonas aeruginosa | 235 |
| 77 | BS3312 | Pseudomonas aeruginosa | 357 |
| 78 | BS3311 | Pseudomonas aeruginosa | 357 |
| 79 | BS3285 | Pseudomonas aeruginosa | NA |
| 80 | BS3284 | Pseudomonas aeruginosa | 235 |
| 81 | BS3283 | Pseudomonas aeruginosa | 357 |
| 82 | BS3282 | Pseudomonas aeruginosa | 357 |
| 83 | BS3281 | Pseudomonas aeruginosa | 235 |
| 84 | BS3280 | Pseudomonas aeruginosa | 357 |
| 85 | BS3279 | Pseudomonas aeruginosa | 357 |



Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 86 | BS3278 | Pseudomonas aeruginosa | 244 |
| 87 | BS3277 | Pseudomonas aeruginosa | 446 |
| 88 | BS3276 | Pseudomonas aeruginosa | 357 |
| 89 | BS3275 | Pseudomonas aeruginosa | 1076 |
| 90 | BS3274 | Pseudomonas aeruginosa | 357 |
| 91 | BS3273 | Pseudomonas aeruginosa | 235 |
| 92 | BS3272 | Pseudomonas aeruginosa | 235 |
| 93 | BS3271 | Pseudomonas aeruginosa | 357 |
| 94 | BS3270 | Pseudomonas aeruginosa | 357 |
| 95 | BS3269 | Pseudomonas aeruginosa | 357 |
| 96 | BS3268 | Pseudomonas aeruginosa | 235 |
| 97 | BS3267 | Pseudomonas aeruginosa | 357 |
| 98 | BS3266 | Pseudomonas aeruginosa | 357 |
| 99 | BS3257 | Pseudomonas aeruginosa | 357 |
| 100 | BS3256 | Pseudomonas aeruginosa | 357 |
| 101 | BS3255 | Pseudomonas aeruginosa | 235 |
| 102 | BS3254 | Pseudomonas aeruginosa | 1076 |

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 103 | BS3253 | Pseudomonas aeruginosa | 1076 |
| 104 | BS3252 | Pseudomonas aeruginosa | 1076 |
| 105 | BS3251 | Pseudomonas aeruginosa | 446 |
| 106 | BS3250 | Pseudomonas aeruginosa | 357 |
| 107 | BS3249 | Pseudomonas aeruginosa | 357 |
| 108 | BS3248 | Pseudomonas aeruginosa | 357 |
| 109 | BS3247 | Pseudomonas aeruginosa | 357 |
| 110 | BS3246 | Pseudomonas aeruginosa | 357 |
| 111 | BS3245 | Pseudomonas aeruginosa | 823 |
| 112 | BS3244 | Pseudomonas aeruginosa | 455 |
| 113 | BS3243 | Pseudomonas aeruginosa | 455 |
| 114 | BS3242 | Pseudomonas aeruginosa | 455 |
| 115 | BS3241 | Pseudomonas aeruginosa | 823 |
| 116 | BS3240 | Pseudomonas aeruginosa | 823 |
| 117 | BS3230 | Pseudomonas aeruginosa | NA |
| 118 | BS2379 | Pseudomonas aeruginosa | 823 |
| 119 | BS2376 | Pseudomonas aeruginosa | 446 |



Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 120 | BS2375 | Pseudomonas aeruginosa | 446 |
| 121 | BS2374 | Pseudomonas aeruginosa | 446 |
| 122 | BS2373 | Pseudomonas aeruginosa | 446 |
| 123 | BS2372 | Pseudomonas aeruginosa | 235 |
| 124 | BS2371 | Pseudomonas aeruginosa | 235 |
| 125 | BS2370 | Pseudomonas aeruginosa | 235 |
| 126 | BS2369 | Pseudomonas aeruginosa | 235 |
| 127 | BS2368 | Pseudomonas aeruginosa | 823 |
| 128 | BS2367 | Pseudomonas aeruginosa | 823 |
| 129 | BS2366 | Pseudomonas aeruginosa | 823 |
| 130 | BS2365 | Pseudomonas aeruginosa | 1182 |
| 131 | BS2364 | Pseudomonas aeruginosa | 446 |
| 132 | BS2363 | Pseudomonas aeruginosa | 823 |
| 133 | BS2361 | Pseudomonas aeruginosa | 823 |
| 134 | BS2360 | Pseudomonas aeruginosa | 357 |
| 135 | BS2359 | Pseudomonas aeruginosa | 823 |
| 136 | BS2358 | Pseudomonas aeruginosa | 823 |



Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 137 | BS2355 | Pseudomonas aeruginosa | 823 |
| 138 | BS2354 | Pseudomonas aeruginosa | 357 |
| 139 | BS2352 | Pseudomonas aeruginosa | 357 |
| 140 | BS2351 | Pseudomonas aeruginosa | 357 |
| 141 | BS2350 | Pseudomonas aeruginosa | 357 |
| 142 | BS2349 | Pseudomonas aeruginosa | 357 |
| 143 | BS2348 | Pseudomonas aeruginosa | 357 |
| 144 | BS2347 | Pseudomonas aeruginosa | 357 |
| 145 | BS2346 | Pseudomonas aeruginosa | 235 |
| 146 | BS2345 | Pseudomonas aeruginosa | 823 |
| 147 | BS2344 | Pseudomonas aeruginosa | 823 |
| 148 | BS2343 | Pseudomonas aeruginosa | 823 |
| 149 | BS2342 | Pseudomonas aeruginosa | 823 |
| 150 | BS2341 | Pseudomonas aeruginosa | 823 |
| 151 | BS2340 | Pseudomonas aeruginosa | 823 |
| 152 | BS2260 | Pseudomonas aeruginosa | 823 |
| 153 | BS2259 | Pseudomonas aeruginosa | 235 |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 154 | BS2258 | Pseudomonas aeruginosa | 823 |
| 155 | BS2257 | Pseudomonas aeruginosa | 823 |
| 156 | BS2256 | Pseudomonas aeruginosa | 823 |
| 157 | BS2255 | Pseudomonas aeruginosa | 823 |
| 158 | BS2254 | Pseudomonas aeruginosa | 823 |
| 159 | BS2253 | Pseudomonas aeruginosa | 823 |
| 160 | BS2252 | Pseudomonas aeruginosa | 823 |
| 161 | BS2251 | Pseudomonas aeruginosa | 823 |
| 162 | BS2250 | Pseudomonas aeruginosa | 823 |
| 163 | BS2249 | Pseudomonas aeruginosa | 823 |
| 164 | BS2248 | Pseudomonas aeruginosa | 823 |
| 165 | BS2247 | Pseudomonas aeruginosa | 823 |
| 166 | BS2246 | Pseudomonas aeruginosa | 823 |
| 167 | BS2244 | Pseudomonas aeruginosa | 253 |
| 168 | BS2242 | Pseudomonas aeruginosa | 823 |
| 169 | BS2241 | Pseudomonas aeruginosa | 235 |
| 170 | BS2240 | Pseudomonas aeruginosa | 235 |

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 171 | BS2239 | Pseudomonas aeruginosa | 235 |
| 172 | BS2238 | Pseudomonas aeruginosa | 1189 |
| 173 | BS2237 | Pseudomonas aeruginosa | 1189 |
| 174 | BS2236 | Pseudomonas aeruginosa | NA |
| 175 | BS2235 | Pseudomonas aeruginosa | 235 |
| 176 | BS2234 | Pseudomonas aeruginosa | 235 |
| 177 | BS2233 | Pseudomonas aeruginosa | 235 |
| 178 | BS2232 | Pseudomonas aeruginosa | 235 |
| 179 | BS2231 | Pseudomonas aeruginosa | 235 |
| 180 | BS2230 | Pseudomonas aeruginosa | 235 |
| 181 | BS2229 | Pseudomonas aeruginosa | 235 |
| 182 | BS2228 | Pseudomonas aeruginosa | 274 |
| 183 | BS2227 | Pseudomonas aeruginosa | 235 |
| 184 | BS2226 | Pseudomonas aeruginosa | 235 |
| 185 | BS2225 | Pseudomonas aeruginosa | 235 |
| 186 | BS2224 | Pseudomonas aeruginosa | 620 |
| 187 | BS2223 | Pseudomonas aeruginosa | NA |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 188 | BS2222 | Pseudomonas aeruginosa | 235 |
| 189 | BS2221 | Pseudomonas aeruginosa | 235 |
| 190 | BS2220 | Pseudomonas aeruginosa | 235 |
| 191 | BS2218 | Pseudomonas aeruginosa | 235 |
| 192 | BS2216 | Pseudomonas aeruginosa | NA |
| 193 | BS2215 | Pseudomonas aeruginosa | 235 |
| 194 | BS2214 | Pseudomonas aeruginosa | 235 |
| 195 | BS2213 | Pseudomonas aeruginosa | 823 |
| 196 | BS2204 | Pseudomonas aeruginosa | 446 |
| 197 | BS2212 | Pseudomonas aeruginosa | 2951 |
| 198 | BS2211 | Pseudomonas aeruginosa | 235 |
| 199 | BS2210 | Pseudomonas aeruginosa | 235 |
| 200 | BS2209 | Pseudomonas aeruginosa | 235 |
| 201 | BS2208 | Pseudomonas aeruginosa | 446 |
| 202 | BS2207 | Pseudomonas aeruginosa | 235 |
| 203 | BS2206 | Pseudomonas aeruginosa | 235 |
| 204 | BS2205 | Pseudomonas aeruginosa | 446 |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 205 | BS2203 | Pseudomonas aeruginosa | 446 |
| 206 | BS2202 | Pseudomonas aeruginosa | 446 |
| 207 | BS2201 | Pseudomonas aeruginosa | 446 |
| 208 | BS2200 | Pseudomonas aeruginosa | 446 |
| 209 | BS2199 | Pseudomonas aeruginosa | 446 |
| 210 | BS2198 | Pseudomonas aeruginosa | 446 |
| 211 | BS2197 | Pseudomonas aeruginosa | 446 |
| 212 | BS2196 | Pseudomonas aeruginosa | 823 |
| 213 | BS2195 | Pseudomonas aeruginosa | 446 |
| 214 | BS2194 | Pseudomonas aeruginosa | 235 |
| 215 | BS2193 | Pseudomonas aeruginosa | 823 |
| 216 | BS2192 | Pseudomonas aeruginosa | 823 |
| 217 | BS2191 | Pseudomonas aeruginosa | 244 |
| 218 | BS2187 | Pseudomonas aeruginosa | 235 |
| 219 | BS2186 | Pseudomonas aeruginosa | 235 |
| 220 | BS2185 | Pseudomonas aeruginosa | 235 |
| 221 | BS2184 | Pseudomonas aeruginosa | 244 |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | Organism | Calculated MLST (Oxford) |
|---|---|---|---|
| 222 | BS2183 | Pseudomonas aeruginosa | 235 |
| 223 | BS2182 | Pseudomonas aeruginosa | 235 |
| 224 | BS2181 | Pseudomonas aeruginosa | 235 |
| 225 | BS2180 | Pseudomonas aeruginosa | 235 |
| 226 | BS2179 | Pseudomonas aeruginosa | 235 |
| 227 | BS2178 | Pseudomonas aeruginosa | 235 |
| 228 | BS2177 | Pseudomonas aeruginosa | 235 |
| 229 | BS2176 | Pseudomonas aeruginosa | 357 |
| 230 | BS2175 | Pseudomonas aeruginosa | 235 |
| 231 | BS2174 | Pseudomonas aeruginosa | 235 |
| 232 | BS2173 | Pseudomonas aeruginosa | 235 |
| 233 | BS2172 | Pseudomonas aeruginosa | 823 |
| 234 | BS2170 | Pseudomonas aeruginosa | 357 |
| 235 | BS2169 | Pseudomonas aeruginosa | 235 |

Quality check summary report calculated using EPISEQ® CS for the dataset of *Pseudomonas aeruginosa* strains .

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

QC

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 1 | BS3414 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 2 | BS3413 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ⚠ | ✓ |
| 3 | BS3412 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ⚠ | ✓ |
| 4 | BS3411 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 5 | BS3410 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 6 | BS3409 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 7 | BS3407 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 8 | BS3390 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 9 | BS3389 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 10 | BS3388 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 11 | BS3387 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 12 | BS3386 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 13 | BS3385 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 14 | BS3384 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 15 | BS3383 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ✓ |
| 16 | BS3382 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ✓ |
| 17 | BS3381 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 18 | BS3380 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ✓ |
| 19 | BS3379 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ✓ |
| 20 | BS3378 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 21 | BS3377 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 22 | BS3376 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 23 | BS3375 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 24 | BS3374 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 25 | BS3373 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 26 | BS3372 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 27 | BS3371 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 28 | BS3370 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 29 | BS3369 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 30 | BS3368 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 31 | BS3367 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 32 | BS3366 | ⚠ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ | ✓ |
| 33 | BS3365 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 34 | BS3364 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 35 | BS3363 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 36 | BS3362 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 37 | BS3361 | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 38 | BS3360 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 39 | BS3359 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 40 | BS3358 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 41 | BS3356 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 42 | BS3355 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 43 | BS3354 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 44 | BS3350 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 45 | BS3349 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 46 | BS3346 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 47 | BS3345 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 48 | BS3344 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 49 | BS3341 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 50 | BS3340 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 51 | BS3339 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 52 | BS3338 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 53 | BS3337 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 54 | BS3336 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 55 | BS3335 | ⚠ | ✓ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ |
| 56 | BS3334 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 57 | BS3333 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 58 | BS3332 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 59 | BS3331 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 60 | BS3330 | ⚠ | ✓ | ⚠ | ⚠ | ✓ | ⚠ | ✓ |
| 61 | BS3329 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 62 | BS3328 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ⚠ | ✓ |
| 63 | BS3327 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ⚠ | ✓ |
| 64 | BS3326 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 65 | BS3325 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 66 | BS3324 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 67 | BS3323 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ⚠ | ✓ |
| 68 | BS3322 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 69 | BS3321 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 70 | BS3320 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 71 | BS3319 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 72 | BS3317 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 73 | BS3316 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 74 | BS3315 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 75 | BS3314 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 76 | BS3313 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 77 | BS3312 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 78 | BS3311 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 79 | BS3285 | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 80 | BS3284 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 81 | BS3283 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 82 | BS3282 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 83 | BS3281 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 84 | BS3280 | ⚠ | ⚠ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 85 | BS3279 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 86 | BS3278 | ⚠ | ⚠ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 87 | BS3277 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 88 | BS3276 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 89 | BS3275 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 90 | BS3274 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 91 | BS3273 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 92 | BS3272 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 93 | BS3271 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 94 | BS3270 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 95 | BS3269 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 96 | BS3268 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 97 | BS3267 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 98 | BS3266 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 99 | BS3257 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 100 | BS3256 | ⚠ | ⚠ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 101 | BS3255 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 102 | BS3254 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 103 | BS3253 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 104 | BS3252 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 105 | BS3251 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 106 | BS3250 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 107 | BS3249 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 108 | BS3248 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 109 | BS3247 | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ | ⚠ |
| 110 | BS3246 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 111 | BS3245 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 112 | BS3244 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 113 | BS3243 | ⚠ | ⚠ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 114 | BS3242 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 115 | BS3241 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 116 | BS3240 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 117 | BS3230 | ⚠ | ✓ | ⚠ | ✓ | ✓ | ✓ | ✓ |
| 118 | BS2379 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 119 | BS2376 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 120 | BS2375 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 121 | BS2374 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 122 | BS2373 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 123 | BS2372 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 124 | BS2371 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 125 | BS2370 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 126 | BS2369 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 127 | BS2368 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 128 | BS2367 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 129 | BS2366 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 130 | BS2365 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 131 | BS2364 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 132 | BS2363 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 133 | BS2361 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 134 | BS2360 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 135 | BS2359 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 136 | BS2358 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 137 | BS2355 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 138 | BS2354 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 139 | BS2352 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 140 | BS2351 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 141 | BS2350 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 142 | BS2349 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 143 | BS2348 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 144 | BS2347 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 145 | BS2346 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 146 | BS2345 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 147 | BS2344 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 148 | BS2343 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 149 | BS2342 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 150 | BS2341 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 151 | BS2340 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 152 | BS2260 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 153 | BS2259 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 154 | BS2258 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 155 | BS2257 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 156 | BS2256 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 157 | BS2255 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 158 | BS2254 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 159 | BS2253 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 160 | BS2252 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 161 | BS2251 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 162 | BS2250 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 163 | BS2249 | ⚠ | ✓ | ✓ | ⚠ | ✓ | ⚠ | ✓ |
| 164 | BS2248 | ⚠ | ✓ | ✓ | ⚠ | ✓ | ⚠ | ✓ |
| 165 | BS2247 | ⚠ | ✓ | ✓ | ⚠ | ✓ | ⚠ | ✓ |
| 166 | BS2246 | ⚠ | ✓ | ✓ | ⚠ | ✓ | ⚠ | ✓ |
| 167 | BS2244 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 168 | BS2242 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 169 | BS2241 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 170 | BS2240 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 171 | BS2239 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 172 | BS2238 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 173 | BS2237 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 174 | BS2236 | ⚠ | ✓ | ✓ | ⚠ | ⚠ | ✓ | ⚠ |
| 175 | BS2235 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 176 | BS2234 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 177 | BS2233 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 178 | BS2232 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 179 | BS2231 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 180 | BS2230 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 181 | BS2229 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 182 | BS2228 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 183 | BS2227 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 184 | BS2226 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 185 | BS2225 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 186 | BS2224 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 187 | BS2223 | ⚠ | ⚠ | ⚠ | ⚠ | ✓ | ⚠ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 171 | BS2239 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 172 | BS2238 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 173 | BS2237 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 174 | BS2236 | ⚠ | ✓ | ✓ | ⚠ | ⚠ | ✓ | ⚠ |
| 175 | BS2235 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 176 | BS2234 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 177 | BS2233 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 178 | BS2232 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 179 | BS2231 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 180 | BS2230 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 181 | BS2229 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 182 | BS2228 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 183 | BS2227 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 184 | BS2226 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 185 | BS2225 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 186 | BS2224 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 187 | BS2223 | ▲ | ▲ | ▲ | ⚠ | ✓ | ⚠ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 205 | BS2203 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 206 | BS2202 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 207 | BS2201 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 208 | BS2200 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 209 | BS2199 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 210 | BS2198 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 211 | BS2197 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 212 | BS2196 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 213 | BS2195 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 214 | BS2194 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 215 | BS2193 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 216 | BS2192 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 217 | BS2191 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 218 | BS2187 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 219 | BS2186 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 220 | BS2185 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 221 | BS2184 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Report for panel: Phase 1 and 2 (235 samples) - Generated by manisha goyal on 2021-03-01

| # | Strain | QC Summary | Raw data | Trimmed data | Assembly | Identity check | Contamination | Marker calling |
|---|---|---|---|---|---|---|---|---|
| 222 | BS2183 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 223 | BS2182 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 224 | BS2181 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 225 | BS2180 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 226 | BS2179 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 227 | BS2178 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 228 | BS2177 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 229 | BS2176 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 230 | BS2175 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 231 | BS2174 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 232 | BS2173 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 233 | BS2172 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 234 | BS2170 | ⚠ | ✓ | ✓ | ✓ | ✓ | ⚠ | ✓ |
| 235 | BS2169 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Legend: ✓ OK ⚠ Minor warning ▲ Major warning

## References


BABALOLA, O. O. 2003. Molecular techniques: An overview of methods for the detection of bacteria. *African journal of biotechnology,* 2**,** 710-713.

DELLINGER, E. P. 2016. Prevention of hospital-acquired infections. *Surgical infections,* 17**,** 422-426.

DYLUS, D., PILLONEL, T., OPOTA, O., WÜTHRICH, D., SETH-SMITH, H., EGLI, A., LEO, S., LAZAREVIC, V., SCHRENZEL, J. & LAURENT, S. 2020. NGS-based S. aureus typing and outbreak analysis in clinical microbiology laboratories: lessons learned from a Swiss-wide proficiency test. *Frontiers in microbiology***,** 2822.

GENOVESE, C., LA FAUCI, V., D'AMATO, S., SQUERI, A., ANZALONE, C., COSTA, G. B., FEDELE, F. & SQUERI, R. 2020. Molecular epidemiology of antimicrobial resistant microorganisms in the 21th century: a review of the literature. *Acta Bio Medica: Atenei Parmensis,* 91**,** 256.

GOLDBERG, B., SICHTIG, H., GEYER, C., LEDEBOER, N. & WEINSTOCK, G. M. 2015. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio,* 6**,** e01888-15.

GOYAL, M., HAUBEN, L., POUSEELE, H., JAILLARD, M., DE BRUYNE, K., VAN BELKUM, A. & GOERING, R. 2020. Retrospective Definition of Clostridioides difficile PCR Ribotypes on the Basis of Whole Genome Polymorphisms: A Proof of Principle Study. *Diagnostics,* 10**,** 1078.

HEALTH, U. D. O. & SERVICES, H. 2019. CDC. Antibiotic resistance threats in the United States, 2019. CDC Atlanta: Atlanta, GA, USA.

JIA, B., RAPHENYA, A. R., ALCOCK, B., WAGLECHNER, N., GUO, P., TSANG, K. K., LAGO, B. A., DAVE, B. M., PEREIRA, S. & SHARMA, A. N. 2016. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research***,** gkw1004.

LASKER, B. A. 2002. Evaluation of performance of four genotypic methods for studying the genetic epidemiology of Aspergillus fumigatus isolates. *Journal of clinical microbiology,* 40**,** 2886-2892.

MAGILL, S. S., EDWARDS, J. R., BAMBERG, W., BELDAVS, Z. G., DUMYATI, G., KAINER, M. A., LYNFIELD, R., MALONEY, M., MCALLISTER-HOLLOD, L. & NADLE, J. 2014. Multistate point-prevalence survey of health care–associated infections. *New England Journal of Medicine,* 370**,** 1198-1208.

MANUKUMAR, H. & UMESHA, S. 2017. MALDI-TOF-MS based identification and molecular characterization of food associated methicillin-resistant Staphylococcus aureus. *Scientific reports,* 7**,** 1-16.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y.-J. & CHEN, Z. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380.

PANUNZI, L. G. 2020. sraX: a novel comprehensive resistome analysis tool. *Frontiers in microbiology***,** 52.

PELEGRIN AC, S. Y., GRIFFON A, PALMIERI M, MIRANDE C, KARUNIAWATI A, SEDONO R, ADITIANINGSIH D, GOESSENS WHF, VAN BELKUM A, VERBRUGH HA, KLAASSEN CHW, SEVERIN JA. 2019. High-risk international clones of carbapenem-nonsusceptible Pseudomonas aeruginosa endemic to

Indonesian intensive care units: impact of a multifaceted infection control intervention analyzed at the genomic level. *mBio* 10**,** e02384-19.

QUAINOO, S., COOLEN, J. P., VAN HIJUM, S. A., HUYNEN, M. A., MELCHERS, W. J., VAN SCHAIK, W. & WERTHEIM, H. F. 2017. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clinical microbiology reviews,* 30**,** 1015-1063.

ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M., LEAMON, J. H., JOHNSON, K., MILGREW, M. J. & EDWARDS, M. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475**,** 348-352.

SÁNCHEZ, G. R. 2015. Identification and typing methods for the study of bacterial infections: a brief review and mycobacterial as case of study.

SHOKOOHIZADEH, L. 2016. Molecular methods for bacterial strain typing. *Medical Laboratory Journal,* 10**,** 1-7.

SYDENHAM, T. V., OVERBALLE-PETERSEN, S., HASMAN, H., WEXLER, H., KEMP, M. & JUSTESEN, U. S. 2019. Complete hybrid genome assembly of clinical multidrug-resistant Bacteroides fragilis isolates enables comprehensive identification of antimicrobial-resistance genes and plasmids. *Microbial genomics,* 5.

TSHIBANGU-KABAMBA, E., NGOMA-KISOKO, P. D. J., TUAN, V. P., MATSUMOTO, T., AKADA, J., KIDO, Y., TSHIMPI-WOLA, A., TSHIAMALA-KASHALA, P., AHUKA-MUNDEKE, S. & MUMBA NGOY, D. 2020. Next-generation sequencing of the whole bacterial genome for tracking molecular insight into the broad-spectrum antimicrobial resistance of Helicobacter pylori clinical isolates from the Democratic Republic of Congo. *Microorganisms,* 8**,** 887.

VALOUEV, A., ICHIKAWA, J., TONTHAT, T., STUART, J., RANADE, S., PECKHAM, H., ZENG, K., MALEK, J. A., COSTA, G. & MCKERNAN, K. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research,* 18**,** 1051-1063.

VAN BELKUM, A., TASSIOS, P., DIJKSHOORN, L., HAEGGMAN, S., COOKSON, B., FRY, N., FUSSING, V., GREEN, J., FEIL, E. & GERNER‐SMIDT, P. 2007. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection,* 13**,** 1-46.

VAN GOETHEM, N., DESCAMPS, T., DEVLEESSCHAUWER, B., ROOSENS, N. H., BOON, N. A., VAN OYEN, H. & ROBERT, A. 2019. Status and potential of bacterial genomics for public health practice: a scoping review. *Implementation Science,* 14**,** 1-16.

XAVIER, B. B., MYSARA, M., BOLZAN, M., RIBEIRO-GONÇALVES, B., ALAKO, B. T., HARRISON, P., LAMMENS, C., KUMAR-SINGH, S., GOOSSENS, H. & CARRIÇO, J. A. 2020. BacPipe: a rapid, user-friendly whole-genome sequencing pipeline for clinical diagnostic bacteriology. *Iscience,* 23**,** 100769.

ZANKARI, E., HASMAN, H., COSENTINO, S., VESTERGAARD, M., RASMUSSEN, S., LUND, O., AARESTRUP, F. M. & LARSEN, M. V. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother,* 67**,** 2640-4.

ZEESHAN, F. & RAZZAK, S. 2020. Next generation sequencing and its role in clinical microbiology and molecular epidemiology. *Annals of Jinnah Sindh Medical University,* 6**,** 31-32.

ZIMLICHMAN, E., HENDERSON, D., TAMIR, O., FRANZ, C., SONG, P., YAMIN, C. K., KEOHANE, C., DENHAM, C. R. & BATES, D. W. 2013. Health care–associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA internal medicine,* 173**,** 2039-2046.

# Chapter 7
## Summary

Routine clinical microbiology laboratories are at the receiving end of new diagnostic technologies. Over the past decades nucleic acid amplification technologies, mass spectrometry and, more recently, omics technologies including next generation sequencing have been successfully introduced for increasingly large-scale diagnostic applications. Particularly in the field of microbial infections, diagnostic procedures have to start with an accurate identification of pathogenic microbes. Rapid technological advancements made this increasingly efficient through the successful application of a variety of methodologies, from more classical ones to the most recent technologies for strain detection and further characterization. The introduction in routine diagnostic laboratories of such technologies in combination with the readily available and extensive amounts of clinical patient-related and demographic data has led to a surge in the implementation of new analytical tools for the combined interpretation of both diagnostic laboratory data and patient-oriented information. I will here summarize the diverse methods that are available for interpretation of such large scale diagnostic data and we will summarize the quality of additional tools that will allow the combined interpretation of "big diagnostic data" and the plethora of patient-oriented, environmental and epidemiological clinical data. The ultimate target for such approaches is to streamline and accelerate data management in favor of improved patient care.

Below I will describe the species specific work described in this thesis. This will be done in the context of the use of new technology to help improve diagnostic tests which for the most part are based on classical technology. I intend to extrapolate my findings to more general applicability in the field of routine diagnostic microbiology.

**How *Staphylococcus aureus* evolved during nasal colonization**

*Staphylococcus aureus* can colonize the human vestibulum nasi longitudinally for many years. It is unknown whether and how *S. aureus* adapts to this ecological niche during colonization. We determined the short (1 and 3 months) and mid-term (36 months) genomic evolution of *S. aureus* in natural carriers and artificially colonized volunteers. Multi-locus

sequence typing (MLST) and single nucleotide polymorphism (SNP) analysis based on whole-genome sequencing (WGS) were carried out. Mutation frequencies within resident bacterial populations over time were quantified using core genome SNP counts and pairwise SNP divergence assessment. SNP counts in all naturally colonizing strains varied from 0 to 757 (median 4) within a period of 1-3 months. These strains showed random and independent patterns of pairwise SNP divergence (0 to 44 SNPs, median 7). When the different core genome SNP counts over a period of 3 years were considered, the median SNP count was 4 (range 0–26). Host-specific pairwise SNP divergence for the same period ranged from 9 to 57 SNPs (median 20). During short term artificial colonization the mutation frequency was even lower (0–7 SNPs, median 2) and the pairwise SNP distances were 0 to 5 SNPs (median 2). Quantifying mutation frequencies is important for the longitudinal follow-up of persistent colonization, epidemics of infections and more local outbreak management. Random patterns of pairwise SNP divergence between the strains isolated from single carriers suggested that the WGS of multiple colonies is necessary in this context. Over periods up to 3 years, maximum median core genome SNP counts and SNP divergence for the strains studied were 4 and 20 SNPs or lower. During artificial colonization, where median core genome SNP and pairwise SNP distance scores were 2, there is no early stage selection of different genotypes. Therefore, we suggest an epidemiological cut off value of 20 SNPs as a marker of *S. aureus* strain identity during studies on nasal colonization and also outbreaks of infection.

The SNP Cutoff or thresholds illustrates the genetic relatedness among the strains thereby could play a potential role in management of *S. aureus* outbreaks by excluding the patients harbouring *S. aureus* strains that are unlikely to be part of the same outbreak and to stay focused only on those who will need further epidemiological follow-up. Present study established SNP cutoff values which could also be implemented for the purpose of genomic surveillance to combat future *S. aureus* infection outbreaks and similar studies could also be performed for other bacterial pathogens.

**Novel Typing technique for *Clostridioides difficile***

*Clostridioides difficile* is a significant cause of sometimes severe health care-associated infections. The epidemiological study of *C. difficile* infection (CDI) traditionally involves PCR ribotyping. However, ribotyping will be increasingly replaced by WGS. This implies that WGS types need correlation with classical ribotypes (RTs) in order to perform

retrospective clinical studies if needed. We selected genomes of hyper-virulent *C. difficile* strains of RT001, RT017, RT027, RT078, and RT106 to try and identify new discriminatory markers using in silico ribotyping PCR and De Bruijn graph-based Genome Wide Association Studies (DBGWAS). First, in silico ribotyping PCR was performed using reference primer sequences and 30 *C. difficile* genomes of the five different RTs identified above. Second, discriminatory genomic markers were sought with DBGWAS using a set of 160 independent *C. difficile* genomes (14 ribotypes). RT-specific genomic polymorphisms were annotated and validated for their specificity and sensitivity against a larger dataset of 2425 *C. difficile* genomes covering 132 different RTs. In silico PCR ribotyping was unsuccessful due to non-specific or missing theoretical RT PCR fragments. More successfully, DBGWAS discovered a total of 47 new markers (13 in RT017, 12 in RT078, 9 in RT106, 7 in RT027, and 6 in RT001) with minimum q-values of 0 to $7.40 \times 10{-}5$, indicating excellent marker selectivity. The specificity and sensitivity of individual markers ranged between 0.92 and 1.0 but increased to 1 by combining two of the new markers, hence providing undisputed RT identification based on a single genome sequence. Markers were scattered throughout the *C. difficile* genome in intra- and intergenic regions. We propose here a set of new genomic polymorphisms that efficiently identify five hyper-virulent RTs utilizing WGS data only. Further studies need to show whether this initial proof-of-principle observation can be extended to all 600 existing RTs.

WGS based markers for the identification and characterization of *C. difficile* showed an perfect example of one of the most important application of WGS approaches. Going beyond classical ribotyping based typing method is itself a benchmark move in the field of diagnostics which is a first step of any outbreak management study. Our proposed WGS based strain characterization methodology has a potential of further extension and validation using other *C. difficile* strains as well as other bacterial pathogens.

**Correlation of genomic variations and mortality caused by SARS-CoV-2**

The current COVID-19 pandemic is caused by the SARS-CoV-2 virus for which many variants at the SNP level have now been identified. I show here that different allelic variants among 692 SARS-CoV-2 genome sequences display a statistically significant association with geographic origin ($p < 0.000001$) and COVID-19 case severity ($p = 0.016$). Geographic variation in itself is associated with both case severity and allelic variation

especially in strains from Indian origin ($p < 0.000001$). Using an new alternative bioinformatics approach I was able to confirm that the presence of the D614G mutation correlates with increased case severity in a sample of 127 sequences from a shared geographic origin in the US ($p = 0.018$). While leaving open the question on the pathogenesis mechanism involved, this suggests that in specific geographic locales certain genotypes of the virus are more pathogenic than others. I here show that viral genome polymorphisms may have an effect on case severity when other factors are controlled for, but that this effect is swamped out by these other factors when comparing cases across different geographic regions. Also a novel BIONUMERICS SARS-CoV-2 plugin tool implemented the SNP-based haplotype variations in a large set of SARS-CoV-2 genome sequences observed in the present study and defines the SARS-CoV-2 population structure and dynamics associated with clinical findings, including fatality rates among patients. Although this study focused on certain genotypes of interest, our approach could be adapted easily to novel variants of SARS-CoV-2 in order to identify unkown samples. This study presents a potential future scope by investigating relationships of different genotypes, viral load and patient outcome to reach out to the actual mechanism playing role in increased pathogenesis.

**First whole genome based analysis of *Pseudomonas aeruginosa***

Carbapenem non-susceptible *Pseudomonas aeruginosa* (CNPA) strains from intensive care units (ICUs) in a referral hospital in Jakarta, Indonesia were recently submitted to detailed epidemiological investigations. It was documented that CNPA transmissions and acquisitions among patients were variable over time and that these were not significantly reduced by a set of infection control measures. Four high risk international CNPA clones (sequence type (ST)235, ST823, ST357, ST446) dominated and carbapenem resistance was due to carbapenemase-encoding genes and mutations in the porin OprD. I here present a more detailed genomic analysis of these four major clones.

With whole genome-based Multi Locus Sequence Typing (wgMLST) of the 4 CNPA clones, three to eleven subgroups with up to 200 allelic variants were observed for each of the CNPA clones. Furthermore, I analyzed the three largest CNPA clone clusters for the presence of wgSNPs to redefine CNPA transmission events during hospitalization. A maximum number 35350 SNPs (including non-informative SNPs) and 398 SNPs (excluding non-informative SNPs) was found in ST235, 34570 SNPs (including non-informative SNPs)

and 111 SNPs (excluding non-informative SNPs) in ST357 and 26443 SNPs (including non-informative SNPs) and 61 SNPs (excluding non-informative SNPs) in ST823. SNPs that are excluding non-informative SNPs were commonly noticed in sensor-response regulator genes. However the majority of non-informative SNPs was found in conserved hypothetical proteins or in uncharacterized proteins. Of note, antibiotic resistance and virulence genes segregated according to the wgSNP analyses. A total of 11 transmission chains for ST235 strains were traceable, followed by 6 and 5 possible transmission chains for ST357 and ST823. The present study demonstrates the value of detailed whole genome sequence analysis for highly refined epidemiological analysis of *P. aeruginosa*. Potentially, similar schemes and approaches can be applied to the epidemiological tracing (both locally but certainly also globally) of any other medically relevant pathogen species, with both microbes and viruses amenable to the same technological approach.

**Evaluation of EPISEQ® CS in comparison with other NGS data analysis pipelines**

NGS is rapidly becoming the new Gold Standard method for the epidemiological tracing of bacterial pathogens. WGS provides a wealth of information on genomic identity of bacterial strains that were isolated from the same clinical environment. This information can be used to define whether or not strains share a common origin. This can have a major impact on nosocomial infection control but in order to do so, the WGS data interpretation should be simplified and made accessible to non-bio-informaticians in an easy and straightforward manner. I here present EPISEQ® CS as a tool for rapid translation of primary WGS data into actionable advice for hospital-based microbiologists and infection control professionals.

Using WGS for *Pseudomonas aeruginosa* as an example, here we carried out preassembly quality assessment of reads, *de novo* genome assembly, comparative strain characterization at the WGS level, Anti-Microbial Resistance (AMR) gene profiling and phylogenetic analysis at a push-button level using EPISEQ® CS. I also compared that with results from bioNumerics. Similar results with a few advantages as well as disadvantages of the two different pipelines were observed. Unlike other available WGS data analysis pipelines EPISEQ® CS works as an automated system for full-blown epidemiological genome analysis, does not require bio-informatic expertise and provides a fully consolidated

output report. However, some additional parametric accesses would be a plus in order to improve the quality and efficiency of EPISEQ® CS.

**Overall message in the THESIS**

Different aspects of whole genome based applications in the field of diagnostics have shown in my thesis. Technological shift from laboratory oriented cumbersome methods to in-silico based approaches can be seen in the present work. However the combination of molecular and in-silico approaches has its own potential and reliability. Whole genome based strain characterization and wgSNP based identification of probable transmission events illustrate the fact that the WGS not only saves time but also help healthcare sectors during severe outbreak management and surviellence. I also included the contribution of promising automated technologies like EPISEQ® CS which has made an easy access to various epidemiological tools at one place that to whithout possessing any expertise to run the tool. However one has to have a deep knowledge of genomics to interpret the results and pick out the desired piece of outcome as per the intital focus of the study.

To a large extent, my thesis has a full focus on the addition of WGS approaches to routine clinical microbiological diagnostics. I have discussed the examples I worked on above and I hope it is clear that all chapters in the current thesis point towards successful application of WGS analysis for answering a diversity of diagnostic questions. The methods, although still too expensive, too laborious and too often requiring expertise that is readily available in high-throughput microbiological testing laboratories, are ready for successful importation into such laboratories. All that is needed is a little more time to make the technology more affordable, a little more rapid and easy to handle. Such developments have been at the heart of WGS for the past decades and this will continue for years to come. I am sure this will increase the quality of microbiological testing and with that the level of care for patients with invasive infections.

My thesis has also reviewed the inevitable need of better management of the steadily growing data repository along with molecular biological advancements. Different microbiological techniques available for microbial detection and characterization were summarized along with the advantage of various roles played by NGS data analysis in the field of epidemiology and infection outbreak management is also presented thourougly in the

thesis. Importance of curated databases for various studies and to promote potential role of data management is the back bone of my work. Also, to provide correct directives to manage sudden outbreaks such as COVID-19, genomic variation based findings in the thesis showed remarkable achievements in the field of diagnostics. However alarming growth of bioinformatics data highlighted the real problem which needs a shift from just cloud storage to an integrated and standardized cloud computing solutions to perform analysis using powerful computational algorithms onto the cloud itself.

**Acknowledgements**

I would like to thank my esteemed supervisor – Dr. Prof. Alex Van Belkum for his invaluable supervision, support and tutelage during the course of my PhD degree. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. My gratitude extends to my collegues at bioMérieux: Sylvain Orenga, Martine Olleon, Maryse Guichard, Isabell Epart, Regine Cuizat and Laurence Devigne for helping me settle down at workplace as well as in the society. I would like to thank my teammates: Magali Dancette, Pierre Mahe, Mattia Palmieri, Andreu Colleo Pelegrin and Gaêl, they have really been very influential in shaping my experimental methods and help me learn new skills. I would like to acknowledge the inevitable role of the ViBrANT consortium for funding my PhD, providing various trainings to develop my academic ardour and interpersonal skills and thanks to all the co-supervisors from ViBrant for their support and guidance to address all the isuues realted to my PhD from time to time. Additionally, I also thank Dr. Andreas Peschel, Eberhard Karls University of Tübingen, Germany for their mentorship. I would like to express my soulful gratitude to the almighty Universe for this wonderful life changing opportunity. At last my deepest appreciation goes out to my parents, my husband and my family for their encouragement, trust and support all through my studies. I also acknowledge the moral support of my friends: Chantal Bruzon and Rashmi Parihar. Although the list of people supported me through my PhD jouney is a very long but I am highly grateful to everyone who so ever played even a tiny role in my life, during this period. Without critics one can never progress, so I also appreciate those who helped me become a refined version of myself with their valuable criticism in any manner.