Year : 2018

Evaluation of next génération sequencing for epidemioiogical investigation of nosocomial pathogens

Gomes Magalhães Barbara

Gomes Magalhães Barbara, 2018, Evaluation of next génération sequencing for epidemioiogical investigation of nosocomial pathogens

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_EC48DDB077814

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Service of Hospital Preventive Medicine

Lausanne University Hospital

Evaluation of next generation sequencing for epidemiological investigation of nosocomial pathogens

Doctoral thesis in Life Sciences (PhD)

Presented to the Faculty of Biology and Medicine
University of Lausanne
by

Bárbara GOMES MAGALHÃES

Master in Microbiology by the University of Aveiro, Portugal

Jury

Prof. Matthias Cavassini, President

P.D. Dr. Dominique S. Blanc, Thesis Director

Prof. Didier Hocquet, expert

Prof. Alex Van Belkum, expert

Lausanne 2018

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Service of Hospital Preventive Medicine

Lausanne University Hospital

Evaluation of next generation sequencing for epidemiological investigation of nosocomial pathogens

Doctoral thesis in Life Sciences (PhD)

Presented to the Faculty of Biology and Medicine
University of Lausanne
by

Bárbara GOMES MAGALHÃES

Master in Microbiology by the University of Aveiro, Portugal

Jury

Prof. Matthias Cavassini, President

P.D. Dr. Dominique S. Blanc, Thesis Director

Prof. Didier Hocquet, expert

Prof. Alex Van Belkum, expert

Lausanne 2018

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof. Matthias **Cavassini**
Directeur·rice de thèse	Monsieur	Dr Dominique **Blanc**
Experts·es	Monsieur	Prof. Didier **Hocquet**
	Monsieur	Prof. Alex **van Belkum**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Bárbara Gomes Magalhães

Master in Microbiology Universidade de Aveiro, Portugal

intitulée

Evaluation of next generation sequencing for epidemiological investigation of nosocomial pathogens

Lausanne, le 15 février 2019

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Matthias Cavassini

Dedicada à pessoa mais importante da minha vida, a minha mãe.

Acknowledgments

First I would like to thank the jury, the presidents Prof. Jacques Besson and Prof. Matthias Cavassini, and the experts Prof. Didier Hocquet and Prof. Alex Van Belkum, for kindly accepting to be part of this committee and for the time they have devoted to this job.

Doing a PhD thesis was one of the most challenging periods in my academic career, both professionally and personally. Luckily I did not have to go through it alone and several people were directly involved during the entire process.

I would like to express my appreciation to my supervisor Dr. Dominique Blanc for allowing me to enroll in such an enriching experience. During these four years of PhD I encountered several obstacles and his guidance and confidence in my abilities encouraged me to overcome them all.

Thank you to Dr. Mohamed Abdelbary for introducing me to the field of next generation sequencing and bioinformatics, and for the company in this not so big research group. To Dr. Benoit Valot, a very special thank you for guiding me in what I thought to be a dead end for my project and for the amazing “Galette des rois franc-comtoise” I had the pleasure to discover.

I would also like to thank Dr. Laurence Senn for helping me with all the epidemiological data needed for this study, and for her kindness and availability.

To my fellow PhDs at the institute, for those still thriving and those who have already left, I thank you for the lunch and coffee break talks, for all the chocolate and “free” coffee, for the calming sessions before and after important meetings, and for enduring my constant bragging about Portuguese gastronomy. Thank you Sara Vaz for making me feel a bit at home with your humor of “gaja do norte”, and to Sara, Noemie, and Dagmara for always checking on me and including me in the Institute’s events, i.e. parties.

My dear friends, you know I am not the most affective person, but I don’t need a lot of hugs and kisses to show you how much I appreciate all the times you worried about my sanity, all the supportive messages, all the advices, and nonsense talks.

My Uni girls, Mariana, Patty, and Raquel, thank you for always making sure I feel your love and appreciation, all three in a very different but special way. I wish us a lifetime of good food, always around! Cláudia, with you I feel JC will always be in the house and we will always be the two coffee addicts in the exposition room trying to make sense of life. So much time has passed, and gladly nothing has changed.

To all my expat friends, a huge thank you! Thank you, in a way, for making me feel included in a country that is not naturally mine and making me believe that good friends can be found in all corners of the world. Sylvie and Catalina, thank you for your motherly protection and affection. I couldn't have dealt with life so well if it wasn't for you. Nadine, for all we experienced together during these 5 years and all the funny stories we have to tell to our grandchildren, I thank you profoundly.

As they say, "behind every successful woman is a group text hyping her up". I'm not sure about the successful part, but no one can hype me better than my girls Catarina, Inês, and Raquel, and for that I'm forever thankful. Thank you BFF for always, always, being there.

I also want to thank my marvelous family for the unconditional support, for being there even if "there" is not so close, for believing, for understanding, and for being proud. Thank you to my little cousins who are like brothers and a sister to me. Padrinho and madrinha, thank you for the support, injections of culture and always great humor. Ruffus Maximus, um beijinho!

A special thank you to uncle Tino and aunt São for making it possible for me to experience a life abroad, for the company, for making me feel at home and cheering me up when I needed the most. Thank you Tio Tino for playing the role of a second father to perfection.

I am so incredibly proud and have profound admiration for all women in my life, in where my dear mom takes center stage. Mãe, pelo amor incondicional que demonstras todos os dias, por seres incansável na busca do meu bem-estar, por me ensinares valores tão importantes pelos quais vivo devotamente, por viveres as minhas alegrias e as minhas tristezas com mais intensidade que eu própria, e me apoiares em tudo o que determinei para o meu futuro, mesmo que não seja a solução mais feliz, agradeço-te profundamente. Esta tese é inteiramente dedicada a ti.

Abstract

Evaluation of next generation sequencing for epidemiological investigation of nosocomial pathogens

Rapid and accurate typing of pathogens is crucial for effective surveillance and outbreak investigation. Although classical typing methods are still well implemented in clinical microbiology laboratories, whole genome sequencing (WGS) is emerging as a powerful molecular typing tool with considerable power of discrimination between outbreak and non-outbreak isolates. This technique has been used to study the epidemiology of important pathogens, such as *Pseudomonas aeruginosa* and *Staphylococcus aureus*.

An increase in *P. aeruginosa* incidence was observed in the intensive care units (ICUs) of the University Hospital of Lausanne. Double locus sequence typing (DLST) detected the presence of three major genotypes during the study period with different epidemiological behaviours. One of the projects developed during this doctoral thesis aimed to use WGS to further investigate these three DLST types. A standard methodology was defined by incorporating open access bioinformatic methods for SNPs analysis using *P. aeruginosa* PA14 as the reference. Results showed an unexpected high number of SNP differences between isolates suspected to be part of an outbreak. The original methodology was altered by adding additional steps of stricter quality filtering which resulted in a more accurate number of SNP differences found. Using a closer reference to each DLST type gave similar SNP differences to when the adapted methodology was used. Changing specific mapping and site coverage thresholds resulted in minor changes in SNPs between isolates. When a definitive methodology was finally chosen, WGS was able to differentiate between outbreak (< 10 SNPs) and non-outbreak isolates, to confirm suspected epidemiological links, and infer relatedness between isolates/environment that were not epidemiologically linked. Combining DLST with the discriminatory power of WGS efficiently elucidated on the *P. aeruginosa* epidemiology in our ICUs.

Genomic data is mainly exploited by SNP analysis or by gene-by-gene methods. The objective of this doctoral thesis' second project was to assess the performance of these genomic methods by using a previously published ST228 Methicillin-resistant *Staphylococcus aureus* (MRSA) dataset. Original published results were compared to the ones obtained with the whole genome SNPs (wgSNPs) and whole genome MLST (wgMLST) tools implemented in BioNumerics. Clustering of isolates was identical between the three analysis and distances were similar between wgSNPs and wgMLST. The advantages of using the BioNumerics wgMLST tool for real-time outbreak

investigation, i.e. no need for a close reference, high interlaboratory reproducibility, and almost no bioinformatic skills needed, turn this method into a simple and easy alternative to other analysis approaches.

Résumé

Évaluation du séquençage de nouvelle-génération pour l'investigation épidémiologique d'agents pathogènes nosocomiaux

Le typage rapide et précis des agents pathogènes est essentiel pour assurer une surveillance ou une investigation d'épidémie efficaces. Bien que les méthodes classiques de typage soient encore bien utilisées dans les laboratoires de microbiologie clinique, le séquençage du génome entier (whole genome sequencing, WGS) est en train de devenir un puissant outil de typage moléculaire avec un pouvoir considérable de discrimination permettant de différencier les isolats épidémiques et non- épidémiques. Cette technique a été utilisée pour étudier l'épidémiologie d'agents pathogènes importants comme *Pseudomonas aeruginosa* et *Staphylococcus aureus*.

Une augmentation de l'incidence de *P. aeruginosa* a été observée dans les services de soins intensifs (intensive care units, ICUs) du Centre Hospitalier Universitaire Vaudois.

Le typage par la méthode du "Double locus sequence typing" (DLST) a permis de détecter la présence de trois génotypes majeurs pendant la période d'étude qui avaient des comportements épidémiologiques différents. Un des projets développés pendant cette thèse de doctorat a eu pour but d'utiliser le séquençage du génome entier afin d'examiner plus profondément ces trois types de DLST. Une méthodologie standard a été définie en incorporant des méthodes bio-informatique en libre accès pour l'analyse de SNPs utilisant le génome de *P. aeruginosa* PA14 comme référence. Les résultats ont montré un nombre élevé inattendu de différences de "single nucleotide polymorphisms" (SNPs) entre les isolats suspectés de faire partie d'une épidémie. La méthodologie originale a été altérée en utilisant les étapes supplémentaires de filtrage de qualité plus stricts, qui a abouti à un nombre plus précis de différences de SNP trouvées. Utiliser une référence plus proche à chaque type de DLST a donné des différences de SNP semblables à celles trouvées lors de l'utilisation de la méthodologie adaptée. Le changement des seuils spécifique de mapping et de couverture de site ont abouti à des changements mineurs de SNP entre les isolats. Quand une méthodologie définitive a été finalement choisie, le séquençage du génome entier a permis de différencier les isolats de l'épidémie (10 SNPs) de ceux qui ne font pas partie de l'épidemie, a ainsi confirmé des liens épidémiologiques soupçonnés et déduit la liaison entre isolats/environnement qui n'étaient pas lié de façon épidémiologique. En combinant le DLST avec le pouvoir discriminant du WGS, cela nous a permis d'élucider efficacement l'épidémiologie de *P. aeruginosa* dans nos services de soins intensifs.

Les données génomiques sont principalement exploitées par l'analyse de SNP ou par des méthodes de comparaison de gène-à-gène. L'objectif du deuxième projet de cette thèse de

doctorat était d'évaluer la performance de ces méthodes génomiques en utilisant un dataset d'isolats de *S. aureus* résistant à la méticilline (MRSA) du ST228 précédemment publié. Les résultats publiés originaux ont été comparés à ceux obtenus avec les SNPs de génomes entiers (wgSNPs) et la méthode du "whole genome MLST" (wgMLST), des outils disponible dans le programme commercial BioNumerics. Le groupement des isolats étaient identiques entre les trois analyses et les distances étaient semblables entre wgSNPs et wgMLST. Les avantages d'utiliser l'outil wgMLST de BioNumerics pour les enquêtes en temps réel sur les épidémies, est qu'il n'est pas nécessaire d'avoir une référence génétiquement proche, une grande reproductibilité inter laboratoire et presque aucune compétence en bio-informatique, . Cette méthode est donc une alternative simple pour l'investigation d'épidémies.

Résumé large public

Évaluation du séquençage de nouvelle-génération pour l'investigation épidémiologique d'agents pathogènes nosocomiaux

Les infections acquis à l'hôpital, ou nosocomial, affectent approximativement 30% des patients dans les unités de soins intensifs et sont associées à une morbidité et une mortalité substantielles. *Pseudomonas aeruginosa* et *Staphylococcus aureus* sont parmi les pathogènes nosocomiaux les plus fréquement rapportés. Si une augmentation de l'incidence de ces pathogènes est observée, une investigation doit être entreprise pour comprendre cette augmentation et identifier une épidémie potentielle. Cette investigation doit combiner autant les données épidémiologiques que microbiologiques afin d'évaluer les possibles transmissions de patient à patient et/ou la possibilité d'une source commune.

Le typage des souches bactériennes permet de mesurer la similarité génétique entre elles. Si des souches sont génétiquement très similaires, on suspecte fortement qu'elles appartiennent à la même chaîne de transmission. Les méthodes de typage conventionnelles examinent généralement une fraction du matériel génétique de ces bactéries. Récemment, avec le développement du séquençage de nouvelle génération, le génome entier peut être analysé, augmentant fortement le pouvoir discriminant de la méthode.

L'objectif de cette thèse de doctorat était d'évaluer le séquençage complet de génomes pour l'investigation épidémiologique des pathogènes nosocomiaux *P. aeruginosa* et *S. aureus*.
Suite à une augmentation de l'incidence de *P. aeruginosa* aux soins intensifs, le typage moléculaire a mis en évidence 3 groupes principaux de patients qui étaient tous infectés par les mêmes génotypes. Notre avons utilisé le séquençage complet de génome sur les souches de ces 3 groupes pour mieux comprendre s'il y a eu transmissions ou pas. Après une importante phase d'optimisation de la méthode d'analyse des données de séquençage complet de génome, les résultats ont montré que les souches qui étaient génétiquement très similaires avaient effectivement des liens épidémiologiques entre elles (transmissions potentielles). En plus de révéler les avantages évidents de l'utilisation de cette méthode pour les enquêtes d'épidémies, cette étude a attiré l'attention sur l'influence des programmes bio-informatiques et des paramètres utilisés sur les résultats obtenus.

Il existe deux approches différentes pour analyser les données de séquençage de nouvelle génération: l'analyse des mutations ponctuelles (SNP, "single nucleotide polymorphism") ou la comparaison gène par gène (wgMLST, "whole genome Multi Locus Sequence Typing"). Nous avons voulu évaluer si ces deux méthodes produisaient des résultats identiques. Pour ce faire, nous avons utilisé une collection de souches d'une épidémie bien décrite de *S. aureus* résistant à la

méticilline (SARM). Les résultats ont montré des résultats très similaires, confirmant que les deux approches peuvent être utilisées pour l'investigation d'épidémies. D'autre part, l'analyse gène-par-gène, disponible dans le programme commercial BioNumerics, a été simple et rapide sans qu'il soit nécessaire d'acquérir de grandes connaissances en bio-informatique.

En conclusion, cette étude montre comment le séquençage complet de génomes peut être une valeur ajoutée aux méthodes classiques utilisées pour les investigations d'épidémies. Il met également en évidence les limites de ces méthodes ainsi que leurs avantages et inconvénients.

Scientific communications

Publications in peer reviewed papers

Magalhães B., Senn L., and Blanc D.S. High-Quality Complete Genome Sequences of Three *Pseudomonas aeruginosa* Isolates Retrieved from Patients Hospitalized in Intensive Care Units. (*Microbiology Resource Announcements, from the American Society for Microbiology Journal. In press.*)

Magalhães B., Valot B., Abdelbary M. H., Prod'hom G., Greub G., Senn L., and Blanc D.S. Unveiling *Pseudomonas aeruginosa* epidemiology in intensive care units through combination of molecular typing and whole genome sequencing. (*Manuscript in preparation)*

Blanc D. S., **Magalhães B. G.**, Abdelbary M., Prod'hom G., Greub G., Wasserfallen J.B., Genoud P., Zanetti G., Senn L. 2016. Hand soap contamination by *Pseudomonas aeruginosa* in a tertiary care hospital: no evidence of impact on patients. J Hosp Infect 93:63-7.

Oral presentation

Magalhães B., Valot B., Abdelbary M.M.H.,Prod'hom G., Greub G., Senn L., Blanc D.S. Challenges in using whole genome sequencing for *Pseudomonas aeruginosa* investigation. Annual Swiss Society for Microbiology (SSM) meeting. Lausanne, Switzerland. 28-30 2018

Poster presentation

Magalhães B., Valot B., Abdelbary M.M.H., Prod'hom G., Greub G., Senn L., Blanc D.S. Investigation of an increase in *Pseudomonas aeruginosa* incidence in ICUs using whole genome sequencing. 28th European Congress of Clinical Microbiology and Infectious Diseases (ECCMID). Madrid, Spain. 21-24 2018

Magalhães B., Abdelbary M.M.H., Eggimmann P., Prod'hom G., Greub G., Senn L., Blanc D.S. Application of whole genome sequencing to investigate an increase in *Pseudomonas aeruginosa* incidence in ICUs**.** 6th Applied Bioinformatics and Public Health Microbiology conference. Hinxton, Cambridge, UK. 17-19 May 2017

Magalhães B., Abdelbary M.M.H., Tissot F., Basset P., Berger M., Que Y-A, Eggimmann P., Prod'hom G., Greub G., Zanetti G., Senn L., Blanc D.S. Application of Whole Genome sequencing to investigate a two year's *Pseudomonas aeruginosa* outbreak. Joint Annual Meeting of the Swiss Societies for Infectious Diseases (SSI), Hospital Hygiene (SSHH), Tropical Medicine and Parasitology (SSTMP) and the Swiss Specialists for Tropical and Travel Medicine FMH (SSTTM). Montreux, Switzerland. 1-2 September 2016

Magalhães B., Abdelbary M.M.H., Tissot F., Basset P., Berger M., Que Y-A, Eggimmann P., Prod'hom G., Greub G., Zanetti G., Senn L., Blanc D.S. Application of Whole Genome sequencing to investigate a two year's *Pseudomonas aeruginosa* outbreak. 11th International meeting on microbial epidemiological markers (IMMEM XI). Estoril, Portugal. 9-12 March 2016.

Magalhães B., Abdelbary M.M.H., Prod'hom G., Greub G., Wasserfallen J.B., Genoud P., Zanetti G., Senn L., Blanc D.S. Burden evaluation of hand soap's contamination by *Pseudomonas aeruginosa* in a tertiary care hospital using Whole genome Sequencing. 11th International meeting on microbial epidemiological markers (IMMEM XI). Estoril, Portugal. 9-12 March 2016.

Table of contents

List of Figures

List of Tables

CHAPTER 1.
General Introduction

1.1. Typing of nosocomial pathogens

The main role of bacterial typing is to unveil clonal relatedness between different strains within a species (3). Isolates relatedness enables the assessment of the sources and routes of infection, confirms or rules out outbreaks, determines cross-transmission of nosocomial pathogens, recognizes virulent strains and evaluates the effectiveness of the surveillance systems (4, 5). Typing of microorganisms relies on the fact that bacterial genomes are constantly undergoing alterations by genetic mechanisms such as point mutations, recombination, gene loss or acquisition and horizontal gene transfer (6, 7). This genetic diversity within bacterial species leads to the creation of new phenotypes, which may have selective advantages in specific ecological niches (8).

Choosing a molecular typing method will depend on the need of resolution, on the epidemiological context, as well as on the time and geographical scale it is going to be applied (3). The method should have intra- and inter-laboratory reproducibility, interlaboratory portability, and unequivocal interpretation of results, high throughput and appropriateness. In terms of convenience, it must be user-friendly, with low cost, rapid and affordable (9, 10).

1.1.1. Molecular typing methods

For many years, traditional typing strategies based on phenotypes have been applied in clinical microbiology laboratories. The development of molecular typing methods has enabled the introduction of new tools for efficient surveillance and outbreak detection. As a result, more efficient infection control programmes and distribution of resources were implemented across Europe (3). Several molecular typing methods are commonly used to subtype different pathogens, each one with advantages/disadvantages. PCR-based methods with high discriminatory power, such as multiplelocus VNTR fingerprinting (MLVF) (11), can work rapidly in characterizing isolates to contain local outbreaks. If the outbreak has disseminated to various geographical locations, a robust typing method like Pulsed-field gel electrophoresis (PFGE) would be more suitable. Due to its discriminatory power and applicability to different bacteria , PFGE was considered the gold standard method for molecular typing (12). More recent methods, such as multilocus VNTR analysis (MLVA), Single locus sequence typing (SLST) (13), multilocus sequence typing (MLST) (14), SNP or DNA microarray analysis, allow the typing of isolates with a comparable efficiency to PFGE with the advantage that urgent results can be acquired rapidly. Since different typing methods are based on the detection of different genomic target sequences, variations found with one approach may not be detected when applying another typing method. In these cases, combining several different typing techniques can add more precise discrimination of bacterial isolates than using solely one typing approach (3, 15). Whole genome sequencing (WGS) permits a completely unambiguous typing of different bacterial isolates as it can resolve single base differences between two genomes. This confers high resolution to genomic epidemiological investigation and makes WGS a promising ultimate method for bacterial typing. Nonetheless, WGS is still time consuming and expensive in comparison to other conventional typing methods.

1.1.1.1. Double locus sequence typing (DLST)

Double locus sequence typing is a DNA sequence-based method that relies on partial sequencing of two highly variable loci, and it has been successfully used to investigate the epidemiology of *Staphylococcus aureus* and *Pseudomonas aeruginosa* (16-19). Similarly to other sequence-based methods, it gives unambiguous definition of types, allowing inter laboratory comparisons and high reproducibility. In addition, the use of 96-well microtiter plates greatly reduces costs and handling time.

For such reasons, this method can be incorporated into long term routine surveillance programs (16, 17). In the case of *P. aeruginosa*, the two hypervariable loci consist in *ms172* (partial sequencing of 400 base pairs), and *ms217* (350 base pairs) (18). A simple representation of both loci is present in Figure 2. For both loci, an arbitrary number is assigned to each allele that has a distinct sequence. Hence, the final result consists in two numbers that correspond to the DLST type (18).

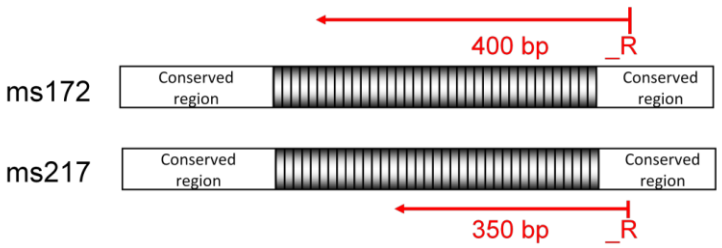


Figure 1. Hypervariable loci *ms172* (400 base pairs) and *ms217* (350 base pairs) used in the typing of *P. aeruginosa* with DLST. (Adapted from (18)).

A study comparing DLST and PFGE showed that, although DLST can be valuable as a first-line typing tool in the investigation of *P. aeruginosa* outbreaks due to its simplicity, its complementation with more discriminatory methods, such as PFGE or WGS, would result in an efficient typing strategy for outbreak investigation (20).

1.1.1.2. Whole genome sequencing (WGS)

In the past decade, DNA sequencing technologies have made important improvements, both quantitatively and qualitatively, increasing accessibility of this technology to research and clinical laboratories worldwide. After the completion of the first human genome sequence (21), different projects aiming to create new cheaper and faster sequencing methods resulted in the development of next generation sequencing (NGS) methods (22).

Advances in the NGS technology resulted in the amelioration of Whole Genome Sequencing (WGS). WGS of bacterial isolates is revolutionizing clinical and public health microbiology with its increased accessibility, decrease sequencing costs, and optimisation of the 'wet laboratory' components of NGS (quality and throughput of DNA extraction, library preparation and sequencing reactions) (23). It enables accurate and rapid species identification, inference of resistome and virulome, and high resolution subtyping without the need for multiple diagnostic steps, which currently involve traditional and molecular methods (24). However, this technology is still far from being universal.

WGS enables a single base-pair resolution between isolates, making it an ultimate molecular typing technique to study bacteria. Sequencing of bacterial genomes is nowadays almost exclusively conducted by Illumina sequencers. Short read sequencing performed with Illumina is based on the principle of *sequencing-by-synthesis*, resulting in read sizes of up to 300 base pairs, and in coverage between 30 and 100 reads per base for a bacterial genome. Longer reads can be produced by other sequencing technologies, such as Pacific Biosciences (PacBio) or Oxford Nanopore's MinION, allowing the complete assembly of bacterial genomes (2). A typical WGS workflow applied to clinical microbiology is represented in Figure 2.

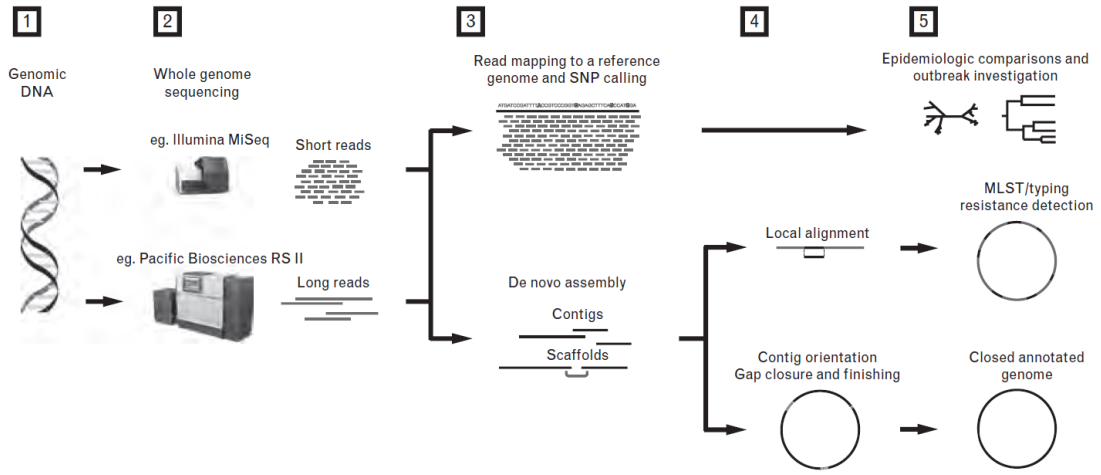


Figure 2.WGS workflow applied to clinical microbiology. 1)DNA extraction from microbial samples. 2) WGS using different next-generation sequencers, Illumina MiSeq and Pacific Biosciences RS II. 3) SNP calling from read mapping to a reference. Reads can also be de novo assembled into longer contiguous sequences (contigs), and orientated and aligned to form scaffolds. 4) The resulting de novo assemblies can be used for further analyses such as typing and resistance detection, or can be further finished into a completed or closed genome. 5) Data analysis for outbreak investigation, typing, or resistance detection. Closed annotated genomes can be used as reference genomes for comparison, or can be analysed in further detai(2).

Nowadays, with advances in obtaining high quality sequencing data, the major problems associated with the implementation of WGS for clinical purposes are focused on post-sequencing data analysis (25). Robust, standardized, portable and scalable methods are needed for the treatment of WGS results in epidemiological investigations. Nevertheless, the existence of an array of bioinformatic tools and approaches for bacterial WGS hinders the harmonization between surveillance and investigation (26). These bioinformatic tools are used to analyse sequencing data with the assistance of computer-based algorithms. Open-source and commercially available bioinformatic programs have been developed for their use in clinical settings by personnel with few knowledge in bioinformatics (27, 28). However, several tools are not able to batch analytical processes on large datasets or customize the analysis pipelines according to the difficulties encountered during the investigation. Bioinformatic software based on text-based command-line in UNIX or Linux operating systems can overcome these disadvantages as

it allows custom programming scripts and pipelines to automate WGS data analysis in a single step (25). When choosing the bioinformatic tools that best suit the WGS analysis approach for the isolate collection, several factors have to be considered, and they are listed in Table 1.

Investigation of outbreak isolates can be performed by calling variants based on analysis of single nucleotide polymorphisms (SNPs), which offers the highest resolution and discrimination, although it is difficult to standardize. Other approaches based on gene-by-gene analysis, referred to as core genome (cg) or whole genome (wg) MLST, may be advantageous as standardization is easier between laboratories using the same scheme. Nevertheless, there is not yet a consensus on the best WGS data analyses methodology, as well as on the cut-offs to determine closely related isolates in an outbreak (2).

1.2. Nosocomial pathogens

1.2.1. Pseudomonas aeruginosa

Pseudomonas aeruginosa is the best-known and studied member of the genus *Pseudomonas*. Members of this genus are gram-negative, rod-shaped, motile bacteria (29) and they possess a remarkable metabolic versatility which allows them to colonize very different ecological niches, from the environment to the interaction with different hosts (30). They are found in water and soil, and on plants, including fruits and vegetables (31). Besides colonizing humans, animals and plants, *P. aeruginosa* is a ubiquitous organism, highly disseminated through the environment, mostly in moist and wet niches (32). A variety of carbon/energy sources are exploited for growth by *P. aeruginosa*, such as carbohydrates, amino acids, fatty acids and, by preference, tricarboxylic acid intermediates (33, 34).

Table 1. Criteria to be considered when choosing bioinformatic tools for WGS data analysis. (Adapted from (25)).

Criteria	Explanation
Usability	Linux-based tools able to easily customize WGS analyses. Tools operated through a Graphical User Interface (GUI) will be preferred by users with little bioinformatics knowledge
Automation	Linux-based pipelines capable of 'batching' or sequential running of several processes on multiple genomes with a single command, compared with running each component individually
Speed	Bioinformatic tools able to analyse multiple samples at the same time and perform computer multithreading or hyperthreading (split large complex processes into smaller processes running in parallel)
Accuracy and detail	While research pursuits require accurate and detailed analyses, the additional resolution from this level of detail is not always required for clinical decisions. For example, Bayesian methods have become popular in estimating a phylogenetic tree. However, while faster neighbour-joining methods may not produce as accurate an evolutionary tree, the resolution is sufficient and rapid analysing a public health outbreak in real time
Cost	Free publicly available software for bioinformatic analysis tends to be command line based with low adaptability across different sequencing platforms. GUI-based software that can be used with relatively little experience is available but at a financial, speed, and sometimes detail, cost
Documentation and support	Commercial software offers user manuals and professional support for troubleshooting. In open-source software, while there is usually some documentation for use but limited support available from open-source software developers, many issues require local computing expertise for implementation and troubleshooting

Additionally, this bacteria can reduce of nitrogen-containing compounds (35). Its ubiquitous growth capacity combined with a high intrinsic resistance against antibiotics and disinfectants, as well as the ability to readily acquire resistance mechanisms makes *P. aeruginosa* an important pathogen for humans (36).

P. aeruginosa population structure is consensually believed to be panmictic-epidemic (37-39), i.e. a superficially clonal structure with frequent recombination that creates new strains with unique genetic characteristics, in which occasionally highly successful epidemic clones arise. In addition, clinical isolates are indistinguishable from environmental isolates; and there are no specific clones related to a specific habitat selection (39).

1.2.1.1. Genome

In 2000, Stover *et al.* published the first complete genome sequence of *P. aeruginosa* (40). This discovery brought new insights on the bacterium as a pathogen, as well as on the relationship between genome size, genetic complexity and ecological versatility. *P. aeruginosa* is currently known to have a very large genome varying from 5.5 to 7 million base pairs that can encode more than 5500 genes. Of this set, more than 500 are involved in gene regulation, allowing the bacterium to switch on/off phenotypes required in specific ecological niches (40, 41).

P. aeruginosa has a mosaic structure consisting of accessory genomic segments inserted in the chromosome at so called “regions of genome plasticity” (RPG) (42). An early comparative genomic study done on five genomes showed that approximately 90 % of the *P. aeruginosa* genome is highly conserved with low sequence diversity (0.5-0.7 %). However, discrepancies are still observed in the core genome size and the genes that it incorporates (43). In combination with deletions, rearrangements and mutations, the horizontal gene transfer of accessory genes plays an important role in the evolution of *P.*

aeruginosa genome. Integrative and conjugative elements (ICEs), replacement islands, prophages and phage-like elements, transposons, insertion sequences, integrons and, in the same strains, extra-chromosomal plasmids compose a great part of the *P. aeruginosa* accessory genome (44). This accessory genome is rich in virulence and in antibiotic resistance genes, which contribute to its importance in healthcare settings (45). Large inversions and recombination events were observed between different *P. aeruginosa* strains, highlighting the high plasticity of *P. aeruginosa* chromosome (42).

1.2.1.2. Pathogenesis and virulence factors

The opportunistic pathogen *P. aeruginosa* disseminates from its reservoirs and infects animals and humans. In the latter case, it can cause infections in both community and hospital settings (39). Community-acquired *P. aeruginosa* infections can cause ulcerative keratitis, external otitis, and skin and soft tissue infections (46). *P. aeruginosa* nosocomial infections are responsible for severe and invasive diseases in critically ill and immunocompromised patients (47). This bacterium is the main cause of hospital-acquired pneumonia in ventilated patients (48). It can cause chronic airway infections in patients with bronchiestasis, chronic obstructive bronchopulmonary disease, and cystic fibrosis (CF) (41, 49). Bacteraemia caused by *P. aeruginosa* can occur in neutropenic patients undergoing chemotherapy (46, 50). This pathogen is considered the third leading cause of nosocomial urinary tract infections (UTIs), which can happen through ascending and descending routes, and usually after catheterization or surgery (51, 52). It is extremely probable that a burned patient, or patients with toxic epidermal necrolysis, will be exposed to *P. aeruginosa* during the healing process, due to its presence in the environment (53).

P. aeruginosa capacity to infect several sites, and its persistence in hostile environments, is enabled by the different virulence factors and regulatory mechanisms

encoded in this pathogen's genome. The most common virulence factors are represented in Figure 3.

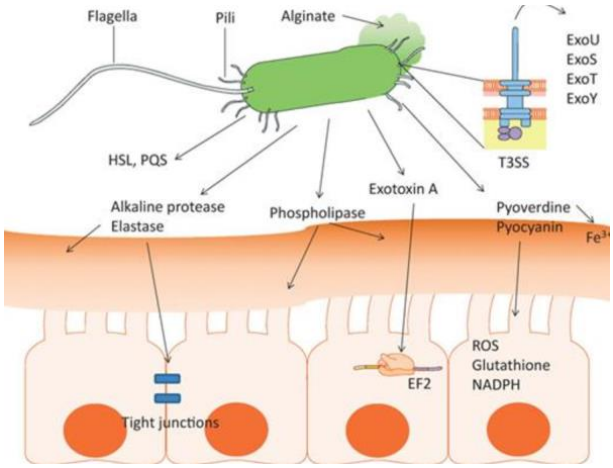


Figure 3. *Pseudomonas aeruginosa* virulence factors (54).

Virulence factors associated with the bacterial surface are flagella, pili, lipopolysaccharides (LPS), components enabling toxin secretion, biofilm formation and quorum sensing (QS) (54, 55). The *P. aeruginosa* cell possesses a single flagellum at the pole where shorter pili are also localized (Figure 3). The flagellum is related to this bacterium swimming motility in aqueous environments, but it is also involved in biofilm dispersal and adhesion to host cells (56, 57). Type IV pili are important adhesins enabling the "twitching motility", which consists of retractile movements that pull bacteria along solid surfaces(58). Additionally, pilli can lead to bacterial aggregation forming microcolonies in specific tissues, and consequently protecting the pathogen from the immune system and antimicrobial treatment (59). LPS is a virulence factor located on the outer bacterial membrane. Its detection generates a strong immune defense, triggering the inflammatory response, exclusion of external molecules, and enabling interactions with antimicrobial agents (60, 61).

Quorum sensing allows bacteria to regulate their population density and their gene expression accordingly (Figure 1). Approximately 6% of the *P. aeruginosa* genome

undergoes regulation by QS systems, which play an important role in biofilm formation and toxin production (62, 63). Biofilms consist of highly organized and structured microbial communities attached to the surface. These communities are encased in extracellular polymeric substances (EPS), which are polymers of polysaccharides, nucleic acids, lipids, and proteins that make up to 90 % of the biofilm volume and confer physical and chemical robustness to the structure (64, 65). Such composition permits the growth of bacteria in a protected mode and allows them to survive in hostile environments (66). *P. aeruginosa* recurs to the type III secretion system (T3SS; Figure 1), a complex pilus-like structure, to translocate effector proteins from bacteria, through bacterial membranes and into the host cytoplasm using a needle-like appendage that forms a pore in the eukaryotic membrane (67). Only four *P. aeruginosa* effector proteins or toxins have been characterized so far: ExoY, ExoS, ExoT, and ExoU (68). The expression of this secretion system usually leads to acute invasive infections, resulting in high mortality rates (69). Secretion of the exopolysaccharide alginate at the cell surface enhances the adhesion capacity and anchor *P. aeruginosa* to the colonized respiratory epithelium, as in the case of respiratory infections (70). Pyocyanin, pyoverdine, alkaline protease, protease IV, elastase, phospholipase and exotoxin A, are other examples of secreted proteins that play an important role in *P. aeruginosa* virulence (61).

1.2.1.3. Antimicrobial resistance

The problem of *P. aeruginosa* resistant strains deserves special attention in many hospitals worldwide, since they are related with a three-fold higher rate of mortality, a nine-fold higher rate of secondary bacteraemia, a two-fold increase in the length of hospital stay, and consequently, a real burden in healthcare costs (71, 72). The complete sequencing of the *P. aeruginosa* PAO1 wild strain brought great knowledge on this microorganism's inherent resistance. Due to this bacteria genome's flexibility,

"pathogenicity islands" are capable of easily acquire large mobile genetic elements that encode for resistance genes (40). *P. aeruginosa* genome complexity and versatility confers it the capacity to resist a wide variety of antimicrobial agents, posing a serious problem in the choice of therapeutic strategies for serious infections (73).

P. aeruginosa is intrinsically resistant to most antibiotics due to specific mechanisms such as low outer membrane permeability, the presence of multidrug efflux transporters, and endogenous antimicrobial inactivation (74, 75). Additionally, resistance to antimicrobials can be increased due to acquisition of inheritable traits. Such acquired resistance occurs via chromosomal mutations and horizontal transfer of genetic elements, such as plasmids, transposons, integrons, prophages, and resistance islands. A third type of resistance is known as adaptive resistance and it depends on the adaptation of the bacteria to a new antibiotic or environmental stimulus (76). A number of triggering factors are responsible for the induction of this type of resistance, including antibiotics, biocides, polyamines, pH, anaerobiosis, cations and carbon sources, as well as mechanisms like biofilm formation and swarming (77).

The main classes of anti-pseudomonal agents include ß-lactams, fluoroquinolones and aminoglycosides (52, 75, 78). Resistance to ß-lactams involves ß-lactamases, chromosomally encoded efflux mechanisms that lead to antibiotic expulsion, and a decrease of porins in the outer membrane, which reduce the uptake of the drug (79-81). Resistance to fluoroquinolones (ciprofloxacin in particular) involves mutations in the DNA gyrase and topoisomerase IV genes (82, 83). Aminoglycosides, as amikacin and tobramycin, are used as treatment for patients with CF suffering pulmonary infections caused by *P. aeruginosa* (84). However, due to acquired aminoglycoside-modifying enzymes, rRNA methylases and endogenous efflux mechanisms, this class of anti-pseudomonal antibiotics is associated with high resistance occurrence (85).

In 2015, the European Antimicrobial Resistance Surveillance Network (EARS-Net) reported the currently active antimicrobial groups against *P. aeruginosa.* The list included some fluoroquinolones (e.g. ciprofloxacin and levofloxacin), aminoglycosides (e.g. gentamicin, tobramycin and amikacin), some beta-lactams (piperacillin + tazobactam, ceftazidime, cefepime, imipenem, doripenem, meropenem and the new ceftolozane-tazobactam) and polymyxins (polymyxin B and colistin) (86). In addition, this report showed that most of the countries had resistance percentages above 10% for all antimicrobial groups under surveillance (piperacillin-tazobactam, ceftazidime, fluoroquinolones, aminoglycosides, and carbapenems), suggesting that antimicrobial resistance in *P. aeruginosa* is common in Europe; the previously observed (2011-2014) decreasing trends for fluoroquinolone and aminoglycoside resistance and increasing trend for piperacillin + tazobactam resistance in *P. aeruginosa* continued in 2015; lastly, 13.7% of the P. aeruginosa isolates (N= 12711) were resistant to at least three antimicrobial groups, and 5.5% were resistant to all five examined antimicrobial groups (86).

1.2.1.4. Molecular typing of *P. aeruginosa*

P. aeruginosa possesses a very complex ecology. For that reason, only powerful typing methods can give insight on the relatedness of strains, and consequently on the routes of colonization and/or infection (87). Different molecular typing methods have been used to investigate this pathogen's epidemiology.

The high discriminatory power of PFGE gave it the connotation of "gold standard" for DNA fingerprinting of many microorganisms, as in the case of *P. aeruginosa* (88-90). However, this method as several disadvantages, like long analysis time, the use of expensive and specialized equipment, and is labor-intensive which make it not the optimal method to be used in a large investigation (90, 91). Multiple-locus variable

number tandem repeat (MLVA), which characterizes each isolate by the number of repeats in several loci, has also been applied in different *P. aeruginosa* typing schemes (92, 93). Nevertheless, the definition of types is ambiguous thus hindering inter-laboratory standardization (94). Multi-locus sequence typing (MLST) relies on partial sequencing of seven genes of the core genome and showed to be efficient in the study of the global population structure of *P. aeruginosa* (95). Several studies on *P. aeruginosa* population genetics were performed using this technique, but it was not discriminatory enough to investigate local epidemiology (96). More recent studies on the *P. aeruginosa* evolution and dissemination in hospital settings have been conducted by recurring to whole genome sequencing (WGS) (97-99).

1.2.1.5. Epidemiology of *P. aeruginosa* in the Intensive Care Units (ICUs)

Pseudomonas aeruginosa accounts for 11 – 14% of nosocomial infections. These values reach even higher percentages, 13 – 23%, when the infection is acquired in ICUs (100, 101). Upon admission in ICUs, approximately 2 – 13% of individuals are colonized with *P. aeruginosa*, and 1% is already infected with this pathogen (102). Although its incidence may vary from unit to unit, and from study to study, *P. aeruginosa* is commonly identified as the most frequent microorganism in burn units, being the cause of a large number of wound infections, bacteraemia and ventilator-associated pneumonia in these units (103, 104).

A general overview of the *P. aeruginosa* epidemiology in the ICU suggests that colonization is a crucial aspect to be taken into account, since it represents the true bacterial load within ICUs (105). Outbreak occurrence in ICUs was thought to be mainly caused by environmental sources. Thus, after implementation of control measures, several studies showed a reduction of outbreaks (106, 107). In addition to environmental

reservoirs, *P. aeruginosa* was also found to be part of the endogenous microbiota of 2.6 to 24% of the hospitalized patients (32, 108).

1.2.1.6. *P. aeruginosa* in the University Hospital of Lausanne

In 1998, a molecular epidemiological investigation on *P. aeruginosa* possible sources and transmission was performed in the University Hospital of Lausanne (109). It reported that the acquisition of *P. aeruginosa* from faucets as an exogenous source was an important cause of infection and colonization in ICU patients, during a nonepidemic period. Infection control measures were implemented and consequently decreased the incidence of *P. aeruginosa* infection and colonization in patients hospitalized in the ICU, showing its efficiency in the presence of an environmental reservoir and patient to patient transmission (110).

Another study conducted in the same hospital investigated whether *P. aeruginosa* infections in ICU patients were due to endogenous or exogenous sources (111). This study covered a longitudinal period of 10 years, from 1998 to 2007, and used the molecular typing techniques PFGE (1998-2004) and DDSL (2007). The authors concluded that the relative contribution of endogenous and exogenous reservoirs to the colonization and infection of ICU patients with this bacterium varies over time.

More recently, an unexplained increase in *P. aeruginosa* incidence was observed in this hospital's ICUs over a two-year period (112). Clinical and environmental isolates retrieved during the study period were typed using DLST. Several DLST types were found among the isolates. The largest cluster, DLST cluster 1-18, comprised the highest number of patients hospitalized mainly in the burn unit during overlapping periods of time. This DLST type was also found in the environment of the hydrotherapy room. In conclusion, the use of a novel molecular typing method, DLST, lead to the identification of the environmental source of a large burn unit outbreak, which was successfully eradicated

after implementation of a continuous surveillance of DLST type 1-18 *P. aeruginosa* in the ICUs

1.2.2. *Staphylococcus aureus*

Staphylococcus aureus is a Gram-positive, coagulase-positive pathogen belonging to the family *Staphylococcaceae.* This bacterium is spherical, approximately 1 μm in diameter and forming grape-like clusters (113). *S. aureus* is considered a commensal bacterium normally present asymptomatically on skin, skin glands, and mucous membranes, as well as on the nasopharynx, throat, and intestinal tract in 30% of humans (113, 114). This colonization facilitates the acquisition of infections, normally by the *S. aureus* strain the affected individual carries as commensal (115). *S. aureus* is the most clinically important staphylococcal species being responsible for a variety of diseases and clinical outcomes (114). Some of the disease manifestations of this pathogen are bloodstream, skin, soft tissue and lower respiratory tract infections, but it can also cause infections related to medical devices and severe deep-seated infections, such as endocarditis and osteomyelitis (116). *S. aureus* is capable of causing disease in diverse physical settings. Clones of *S. aureus* causing both health-care and community acquired infections have emerged in the past years. These clones transport specific traits responsible for *S. aureus* adaptability to diverse environments with different selective pressures (117). Additionally, *S. aureus* colonizes and causes opportunistic infections in a variety of animal species apart from humans, e.g. livestock-associated infections (118). This capacity of adapting to different environmental and anatomical niches in several host species classifies *S. aureus* as an exceptionally versatile pathogen.

Studies on the population structure of *S. aureus,* using techniques such as PFGE or MLST, have demonstrated the high clonality of this bacterium's population (119). Such

findings are consistent with the perspective that *S. aureus* is not naturally transformable, as opposed to other recombining species (120, 121)

1.2.2.1. Genome

Whole genome sequencing was used to investigate the resistance and virulence mechanisms of *S. aureus*. Methicillin-resistant *Staphylococcus aureus* (MRSA) strains N315 and Mu50 were the first staphylococcal genomes to be sequenced (122) followed by a number of other strains (123-125). The *S. aureus* genome is approximately 2.8 Mbp in size and have a relatively low G+C content. Most regions of the staphylococcal genome are well conserved while several blocks demonstrate high variability, probably due to horizontal acquisition of these genomic islands. Integration of these islands must have, at least initially, required DNA recombination (integrase) genes (126). However, it has been reported that variation from point mutation was 15-fold more frequent than recombination, suggesting that the latter is not the major contributor for genetic variation in *S. aureus* (119). The above-mentioned variable blocks normally carry virulence and antibiotic resistance determinants implicated in the development of staphylococcal diseases, such as as prophages, pathogenicity islands, or staphylococcal cassette chromosomes.

1.2.2.2. Pathogenesis and virulence factors

When *S. aureus* is initially exposed to host tissues beyond the mucosal surface or skin, an upregulation of virulence genes occurs (127). On the other hand, host phagocytes and epithelial cells in the skin and mucosal tissue respond to bacterial products or tissue injury by immune system activation. *S. aureus* α-toxin, β-toxin, and PVL are implicated in pneumonia and lung injury. Both α-toxin and Panton-Valentine leukocidin (PVL) produced by *S. aureus* are pore-forming toxins, which exaggerate the host inflammatory

response by inducing the expression of proinflammatory cytokines and lysing inflammatory cells to release additional inflammatory mediators (128). *S. aureus* can overcome opsonisation by complement and antibodies through the expression of a capsule, clumping factor A, protein A, and several complement inhibitors on its surface. All of these will prevent host opsonins from binding or targeting the bacterium for destruction(129).

In addition to host immune defence evasion, bacterial survival within the host relies on the successful acquisition of nutrients, such as iron. *S. aureus* is able to secrete aureochelin and staphyloferrin during iron starvation, which are high affinity iron-binding particles(130) .

Another virulence mechanism of clinical significance is biofilm formation which allows *S. aureus* to persist on plastics and resist host defences or antibiotics (131). Small colony variants aid *S. aureus* to survive in a metabolically inactive state under harsh conditions. This virulence factor has been implicated in chronic infections, e.g. osteomyelitis (132).

Methicillin-resistant *S. aureus* (MRSA) deserves special consideration when discussing *S. aureus* pathogenesis as it possesses a distinct epidemiology particularly marked by morbidity and mortality (129). In 2005, invasive diseases and deaths attributable to MRSA were 94,360 and 18,650, respectively, in the United States, overcoming mortality rates attributed to HIV (133). Hospital- and community-acquired MRSA are two genotypically different groups of MRSA that target different but overlapping populations and cause different diseases. HA-MRSA became increasingly problematic in the 1990's especially in intensive care unit settings where it became a major cause of nosocomial infections (134). This pathogen's chromosome contains large staphylococcal cassettes (SCC*mec* types I-III), which encodes one (SCC*mec* type I) or

multiple antibiotic resistance genes (SCC*mec* type II and III). Such high resistance to antibiotics probably was the cause for this bacterium's survival in an environment where antibiotic use is frequent (129).

1.2.2.3. Molecular typing of *S. aureus*

Phenotypic methods, like phage typing and protein profiling, were used in variation investigations of *S. aureus* populations. Early *Staphylococcus* taxonomists helped to define staphylococcal biotypes but they correlated loosely with their host species association based on phenotypic markers like coagulation of human and bovine plasma, production of fibrinolysin, crystal violet reaction type, beta haemolytic activity, and phage susceptibility(135). From there after, the development of molecular typing techniques provided increasing resolution for distinguishing *S. aureus* isolates and understanding its population structure. With multilocus enzyme electrophoresis (MLEE) it was possible to infer allelic variation among *S. aureus* strains based on electrophoresis of housekeeping enzymes with varying charge(136). PFGE was considered by several authors as the gold standard for outbreak investigation (90). DNA-based molecular approaches such as multilocus sequence typing (MLST) allowed the investigation of genetic diversity between strains of the same species (14). Combination of allelic variants is used to assign a sequence type (ST), and *S. aures* STs that share alleles at ≥5 loci are considered to belong to the same clonal complex (CC) (137). Due to the availability and affordability of DNA sequence technology, several sequenced-based typing methods are now widely used, such as MLST and spa typing (138), which are the most frequently used for *S. aureus*. Characterization of *S. aureus* isolates is now done though combination of different techniques (including the SCCmec type for the characterization of MRSA strains). Nonetheless, the amount of sequencing MLST requires and the high number of primers needed to identify SCCmec types as new types impede the combination of these

methods for clonal characterization due to cost-related reasons. Consequently, SeqNet (http://www.seqnet.org), the European Network of Laboratories for Sequence Based Typing of Microbial Pathogens, suggested spa typing as the primary sequence-based method for determining genetic relatedness of *S. aureus* isolates (137).

Improvement of the WGS technology in recent years has affected the typing of several pathogens, including *S. aureus*. Figure 4 shows the increasing number of studies on general *S. aureus* population biology along with nosocomial investigations using WGS (137)

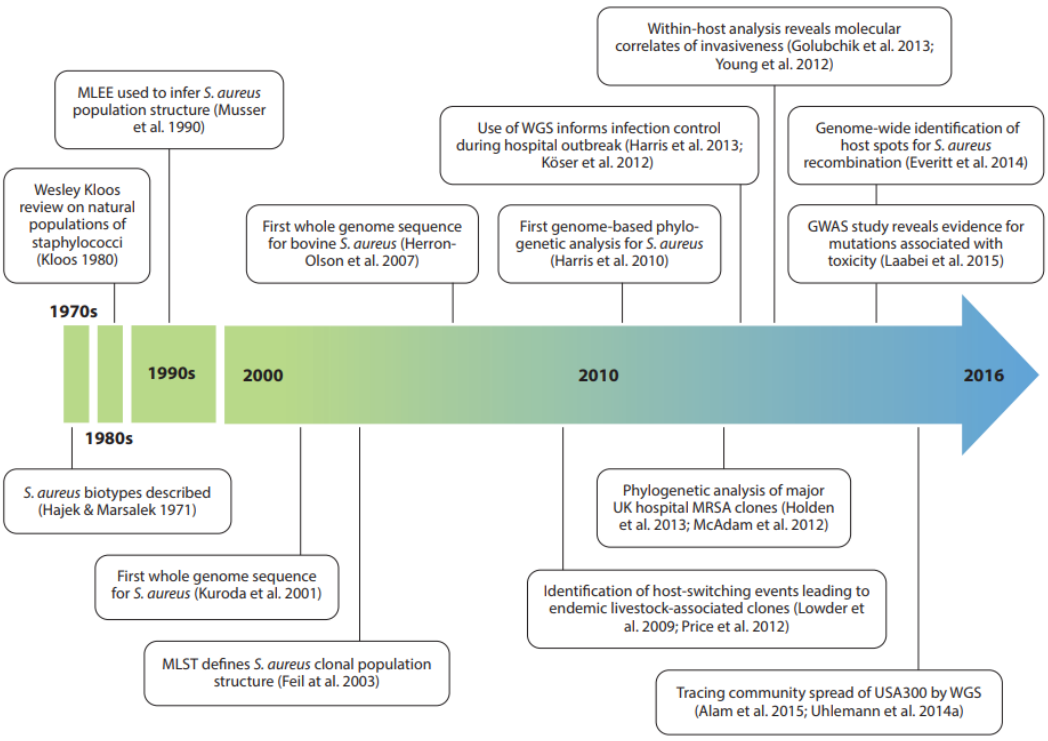


Figure 4. Chronology of different studies on *S. aureus* population biology(137).

CHAPTER 2.

Use of open-access bioinformatic tools to investigate *P. aeruginosa*

2.1. Objectives

An unexplained increase in *P. aeruginosa* incidence was observed in the ICUs of the University Hospital of Lausanne from 2010 to 2014. During this period, the retrieved clinical and environmental isolates were typed by the double locus sequence typing (DLST) method. Numerous DLST clusters were identified, from which three harboured the highest number of patients: cluster 1-18 (N= 24), 6-7 (N= 22), and 1-21 (N= 16). DLST cluster 1-18 isolates were believed to be part of an outbreak in the burn unit (ICU 3), as it was epidemiologically well described before (112). The remaining two DLST types showed sporadic occurrence with only few cases of possible transmission between patients. . The principal objective of this project is to further investigate these three major DLST clusters using the higher discriminatory power of whole genome sequencing. Therefore, the following tasks will be executed:

- Construction of the *P. aeruginosa* isolates phylogeny for each DLST cluster;

- When a definite and accurate phylogenetic tree is achieved, environment-to-patient and patient-to-patient transmission events suspected by epidemiological data will be confirmed/ruled out;
- Genetic characterization of each cluster

2.2. Material and Methods

2.2.1. Bacterial isolates and molecular typing

P. aeruginosa isolates were collected from patients hospitalized in the five ICUs of the University Hospital of Lausanne over a five-year period. From 2010 to 2014, clinical and environmental isolates were typed by the double locus sequence typing (DLST) method previously developed by our group(18). Three major DLST clusters, i.e. clusters with the highest number of patients, were further analysed in this study: DLST cluster 1-18 (24 patients), 6-7 (22 patients), and 1-21 (16 patients). At least one isolate per patient was included. If several isolates were collected from one patient, only isolates sampled 15 days apart were selected, unless they belonged from different sample types. All environmental isolates from the three DLST clusters (mainly from sink traps) were considered. A total of 74 DLST 1-18 isolates (55 clinical and 19 environmental), 50 DLST 6-7 isolates (38 clinical and 12 environmental), and 31 DLST 1-21 isolates (18 clinical and 13 environmental) were selected for whole genome sequencing.

2.2.2. Epidemiological investigation

Epidemiological data (unit and room of hospitalization, dates of ICU admission and discharge, and clinical diagnosis) was retrieved from the hospital databases and used to construct epidemiological maps and annotate the phylogenetic trees. Epidemiological links between patients or environment were identified as: (i) patients hospitalized during

overlapping periods in the same ICU, or (ii) patients showing an identical DLST type with an environmental sample isolated in the same unit during the period of the study.

2.2.3. DNA extraction and whole genome sequencing

We extracted genomic DNA from a 5ml Lysogenic Broth (LB) culture, acquired from single colonies and incubated to reach an early exponential phase, using the GenElute bacterial genomic DNA kit (SIGMA-ALDRICH, St. Louis, MO, USA). Whole genome sequencing was performed on 155 *P. aeruginosa* clinical and environmental isolates by the Lausanne Genomic Technologies Facility (GTF, University of Lausanne). The sequencing libraries were prepared using the Nextera DNA Library Preparation Kit (Illumina, Inc., San Diego, CA, USA) for 100-bp paired-end sequencing runs on Illumina HiSeq 2500, aiming for a 100-fold coverage.

2.2.4. Analysis of WGS data

WGS was performed on 155 clinical and environmental DLST 1-18 (94), DLST 1-21 (31), and DLST 6-7(50) isolates, retrieved from 2004-2014, using Illumina HiSeq. Reads quality metrics were assessed with FastQC. Isolates' sequence type (ST) was assigned from the short reads data by the Short Read Sequence Typing 2 (SRST2) software (139). It defined DLST cluster 1-18, 1-21, and 6-7 as ST1076, ST253, and ST17, respectively. Two methodologies based on mapping raw reads against a reference, SNPs analysis, and phylogeny construction were used in this project. No complete reference genomes belonging to ST1076 or ST17 were published thus far, hence we used a well-known ST253 reference strain, *P. aeruginosa* UCBPP-PA14 (accession number: NC_008463; (140)), for the "mapping against a reference" step in both methodologies. Additionally, the mapping step in the two procedures was performed against a complete reference genome for each ST created by combining both PacBio (Pacific Biosciences) and

Illumina HiSeq sequencing of the index case of each DLST cluster (Patient 2 isolate for DLST 1-18). These three reference genomes were submitted to Microbiology Resource Announcements, from the American Society for Microbiology journal and the genome announcement is currently under revision (Chapter 3).

2.2.4.1. Standard methodology

The bioinformatic analysis of our sequenced data was defined according to a thorough literature search on outbreak investigation of several nosocomial pathogens (141-143). The chosen scheme included essentially open access programs which were responsible each for several steps of the methodology (Figure 5).

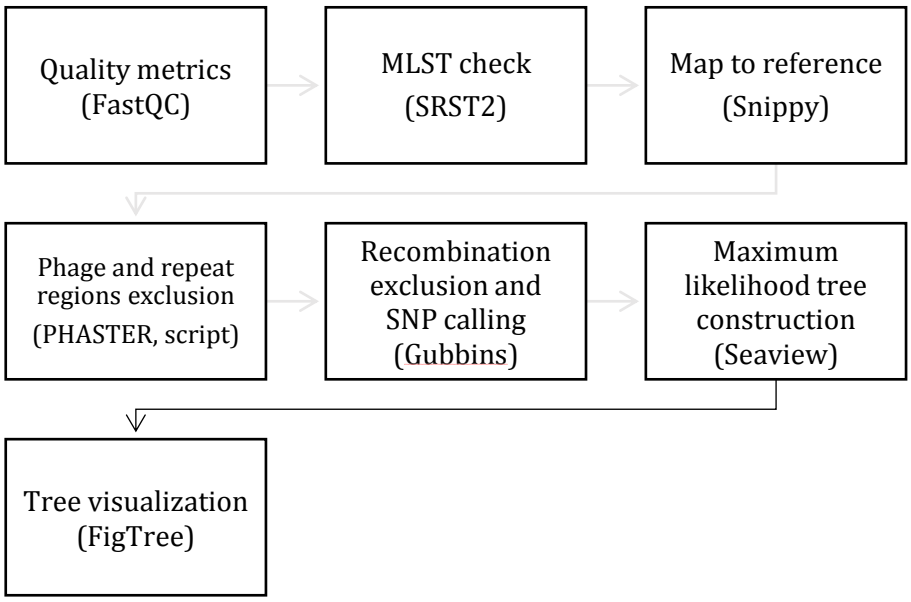


Figure 5. Schematic representation of the different steps included in the standard methodology.

Genome alignment of each DLST cluster isolates was acquired with Snippy (https://github.com/tseemann/snippy) by mapping against the two reference genomes previously mentioned under default parameters (minimum mapping quality of 60, minimum coverage of variant site of 10, and minimum proportion for variant evidence of 0.9). Putative phages present in the reference genomes were searched with PHASTER (144) and repeat regions were identified with a homemade script (developed by Dr.

Benoit Valot, University of Franche-Comté, France) and afterwards masked from the genome alignment. The final single nucleotide polymorphisms (SNPs) alignment was obtained by excluding regions of high SNPs density indicative of recombination. Gubbins (145) was the software used for that purpose by applying the default parameters: minimum of three SNPs to be considered a recombination block, maximum window of 1000, and a minimum window of 100. A maximum likelihood tree was constructed from the final SNPs alignment using the PhyML algorithm implemented in Seaview version 4.7 (146). Tree visualization was done with FigTree version 1.4.3.

2.2.4.2. Adapted methodology

The basis of both methodologies is very similar, yet there are differences in relation to the programs used and the quality filtering applied. A scheme of the second procedure used to analyse the *P. aeruginosa* sequences is present in Figure 6 and the addition of two steps in highlighted in red.

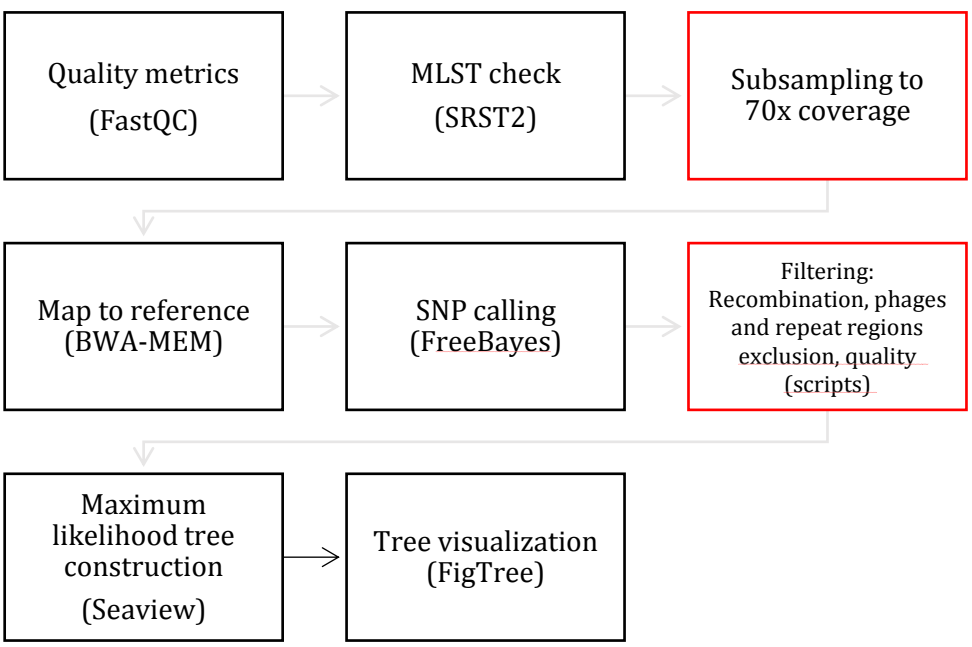


Figure 6. . Schematic representation of the different steps included in the adapted methodology.

A first step of subsampling the number of raw reads to reach the lower read depth observed (70x) was added to provide comparable accuracy in the posterior analysis, as well as to reduce mapping time (147). The subsampled reads were then mapped against their respective complete reference genome with BWA-MEM. Variant calling was performed with FreeBayes with a minimum mapping quality of 60 and a minimum proportion for variant evidence of 0.9. A series of other in-house scripts was applied to the variant call format (VCF) file (lists each position where a SNP is detected along with several characteristics associated with this SNP, e.g. nucleotide change, quality value, or the applied filtering) acquired after SNP calling with FreeBayes. An in-house script was used for identification of recombination regions: it determines a threshold for SNP density according to the data being analysed and lists the regions of high SNP density to be masked above this threshold. A probability to remove regions of high SNP density of 0.001 and a window size of 2000 was used for recombination detection. Additionally, an in-house script performed repeat region identification. Putative phages found with PHASTER (144) along with repeat regions and potential recombination regions were excluded from the genome alignment. The VCF file was then filtered with other in-house scripts applying the following parameter thresholds: minimum quality of base assignement of 100 and a minimum read by allele to report a SNP of 20. A maximum likelihood tree was constructed from the final core SNPs alignment using the PhyML algorithm implemented in Seaview version 4.7 (146). Tree visualization was done with FigTree version 1.4.3.

2.2.4.3. Visualization of SNP differences data

The SNPs sequence alignment in FASTA acquired with both methodologies was converted to a pairwise SNP distance matrix using a script deposited in GitHub (https://github.com/tseemann/snp-dists). A heatmap was constructed from the

pairwise SNP distance matrix using the heatmap.2 function present in the gplots package implemented in R (http://www.R-project.org.). Using the same pairwise matrix as input, the ggplot2 package in R was applied to plot the frequency of the number of SNP differences observed in the dataset.

2.3. Results

2.3.1. Different epidemiology of the three DLST clusters

To infer possible epidemiological links between patients of the same DLST cluster and/or between patients and environmental isolates, the hospitalization period, the ICU where the hospitalization occurred, and the ICUs environmental sampling of *P. aeruginosa* were investigated and are schematically represented in Figure7. DLST cluster 1-18 was previously considered responsible for an outbreak in the burn unit from 2010 to 2012 (112). From the 24 patients harboring this DLST type, 18 were hospitalized in the burn unit (ICU 3), and six in other ICUs. Several epidemiological links were found between the patients hospitalized in the burn unit which shared the same hydrotherapy shower room, during overlapping hospitalization periods. The first patient observed harbouring this DLST type was hospitalized in ICU5 and was not epidemiologically linked to other patients infected with the same type. Links between environmental isolates, mainly from the shower room and sink traps, and patients hospitalized in the same ICU were also found.

Only two epidemiological links were identified for DLST cluster 1-21; one between two patients hospitalized in the same ICU (ICU 2), and one between those patients and an environmental sample retrieved from a sink trap in the same ICU. The remaining patients were dispersed through the six ICUs during the study period, except in 2013 when no patient was found to be colonized or infected with this DLST type (Figure 7). Such behaviour suggests DLST cluster 1-21 was not considered to be the cause of an outbreak.

Three epidemiological links were found between DLST cluster 6-7 patients hospitalized in the burn unit, in 2010. Thereafter, no epidemiological links were suspected as patients were not hospitalized in the same ICU during overlapping periods of time, and no epidemiological links between patients and environmental sources were observed. Similarly to DLST cluster 1-21, this DLST type occurred sporadically throughout the study period and was not responsible for an outbreak.

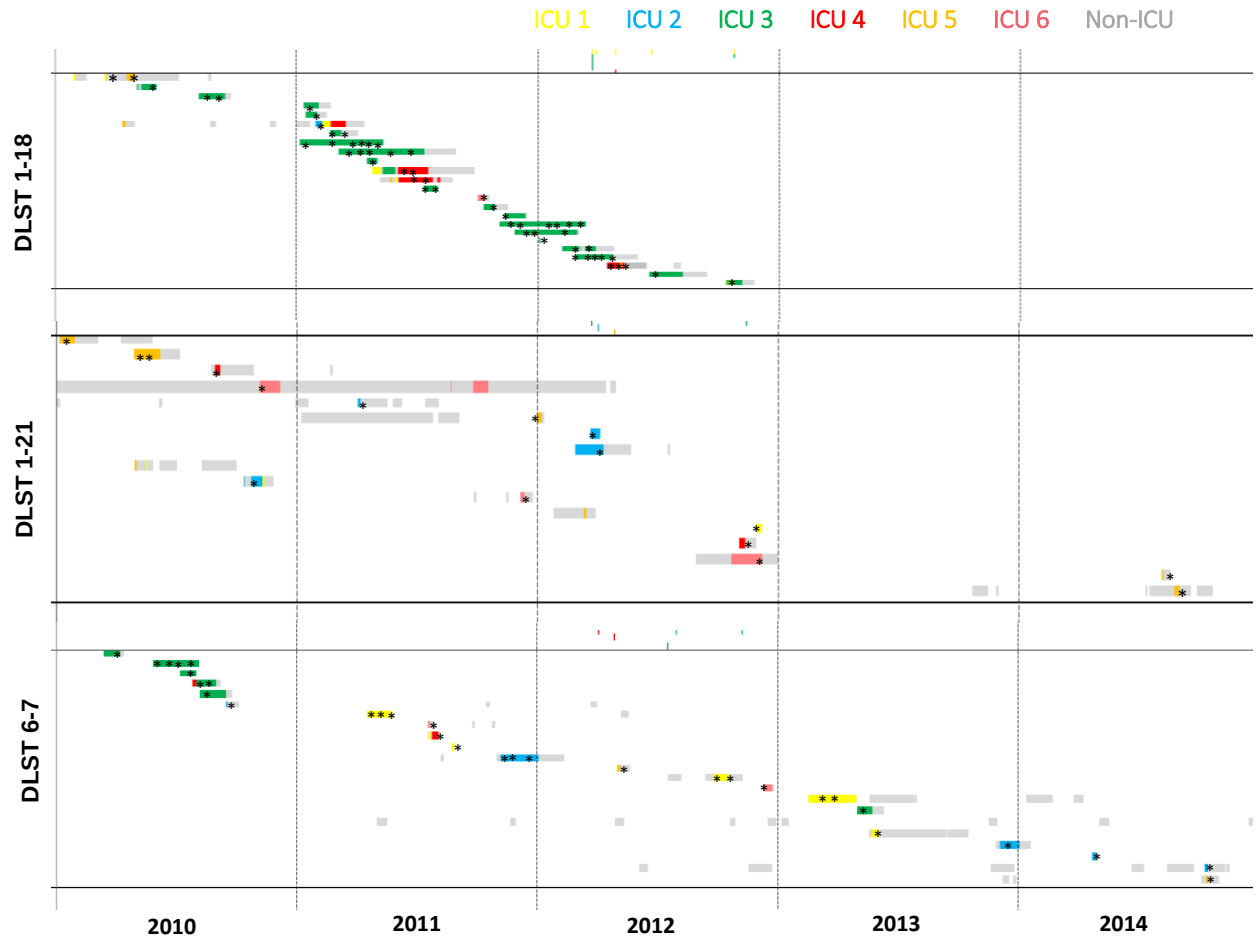


Figure 7. Epidemiological maps of the three different DLST types. The first panel corresponds to patients harbouring DLST 1-18 (N=24), the second to DLST 1-21 (N=16), and the third to DLST 6-7 (N=22). Each line represents the hospitalization period of each patient from 2010 to 2014. Units where patients were hospitalized are differentiated by colors. Stars represent the first isolation of *P. aeruginosa* for each patient.

2.3.2. Different DLST clusters belonged to different sequence types

Among the different genotyes, DLST cluster 1-18, 1-21, and 6-7 comprised the highest number of patients and were chosen for posterior analysis with WGS. Although DLST allows inter laboratory comparison of genotypes (17), the universal standard of MLST is still widely used for strain comparison. Therefore, the Illumina HiSeq raw reads were used to identify the STs present in the isolate collection. MLST results defined DLST 1-18, 1-21, and 6-7 isolates as STs 1076, 253, and 17, respectively.

2.3.3. Standard methodology with mapping against *P. aeruginosa* PA14

A bioinformatic pipeline was created by resorting to open access tools selected according to previously published studies on outbreak investigation using WGS. Although the basic steps of this methodology have been used for WGS data analysis of *P. aeruginosa* isolates, the combination of programs used in this study was only reported for the investigation of other pathogens (2, 148)

DLST 1-18 phylogeny was divided in two clades (definition of clade being a group of all the descendants from a common ancestor): one clade subdivided in two subclades and the other in three, each one with both clinical and environmental isolates (Figure 8). Most of DLST 1-18 isolates were distanced by 20 to 100 SNPs. Patient 1 (Figure 9), which was not hospitalized in the burn unit and had no epidemiological link with the outbreak, clustered within one of the subclades at the tip of a long branch representative of 100 to 200 SNP differences in relation to other isolates. Isolates retrieved from the same patient also exhibited and unexpected high number of SNPs between them. Patient 4 isolates sampled less than one week apart were located in different subclades of the phylogenetic tree (Figure 8, in pink). Three isolates from

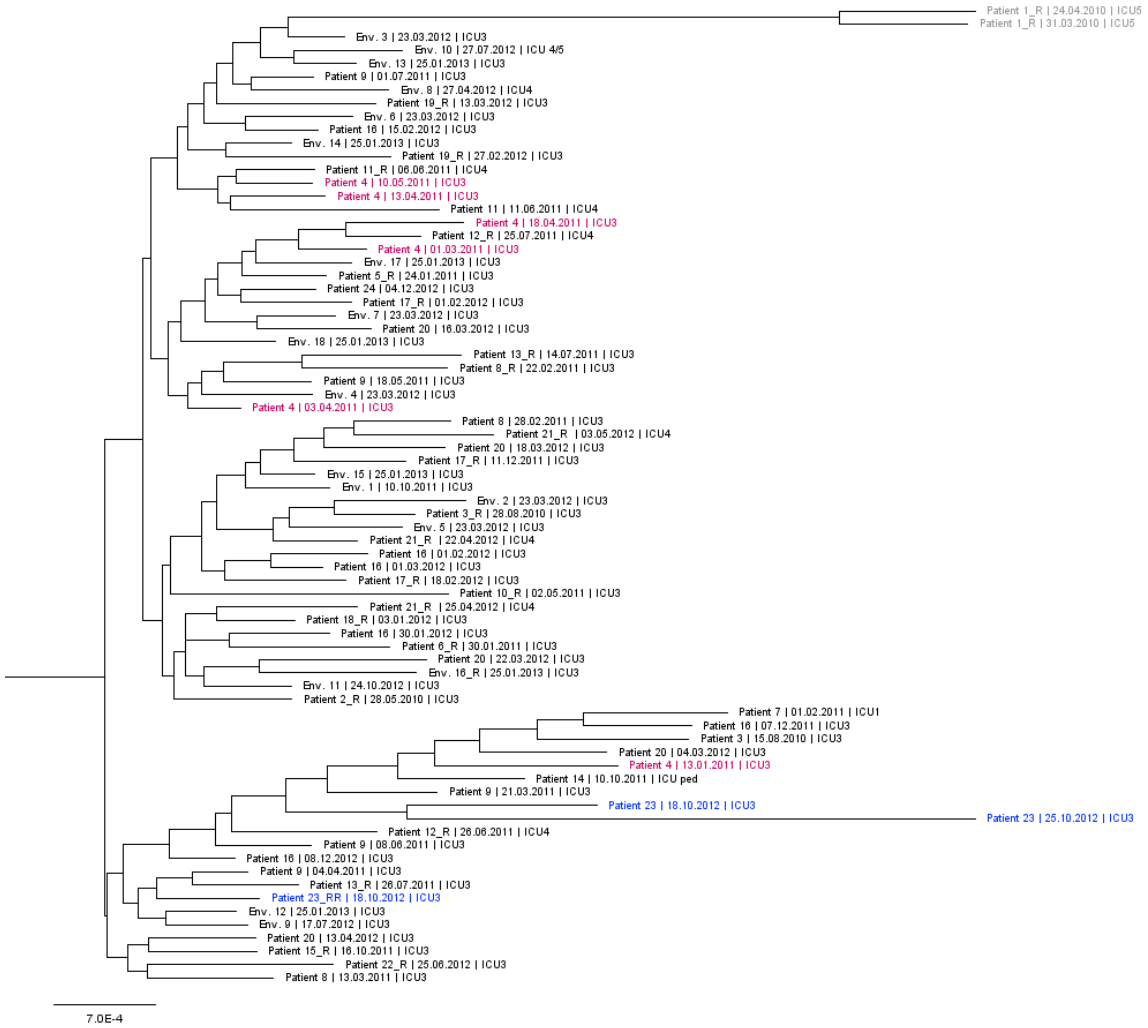


Figure 8. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the standard methodology, mapping against *P. aeruginosa* PA14. Non-outbreak isolates belonging to Patient 1 are highlighted in grey, and clustered apart from the remaining isolates. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

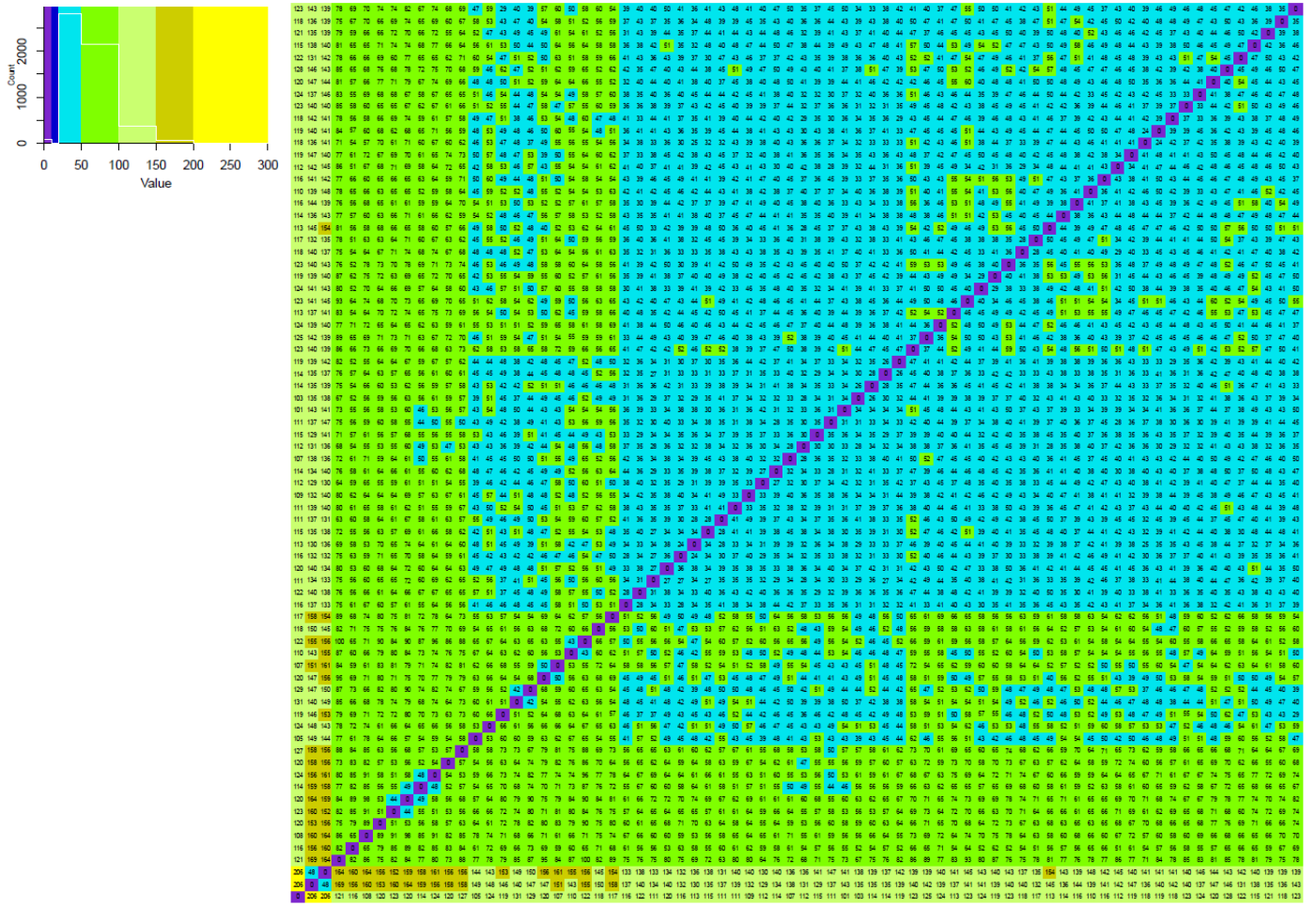


Figure 9. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the standard methodology, mapping against *P. aeruginosa* PA14. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 23 (first isolate) to Patient 21 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 300. A white line on the color legend plot pictures the frequency of each number of SNP differences.

Patient 23 belonged to the same subclade with a high number of SNP differences (80-121 SNPs). Interestingly, two Patient 23 isolates collected seven days apart clustered together with 121 SNPs between them. In spite of being considered responsible for an outbreak, DLST 1-18 demonstrated a minimum number of 24 SNPs between two isolates from different patients and a maximum number of 206. Plotting the frequency of observed SNPs in this isolate collection helped visualize a pique around 40 SNPs, which afterwards decreased to a smaller count of an increasingly higher number of SNPs (Figure 10).

DLST 1-21 phylogeny was divided in numerous clades and subclades (Figure 12) distantly related to each other with a high number of SNPs among them supporting the assumption that this cluster was not responsible for an outbreak (Figure 11). Patient 8 hospitalized in the paediatric ICU clustered apart from the remaining isolates with 81 to 114 SNP differences (Figure 11 and 12). The previously suspected epidemiological link between isolates from three patients and the environment (Figure 12, in blue) was confirmed as no SNP differences were found between them. Two isolates retrieved two years apart, from two patients hospitalized in different ICUs, shared only nine SNP differences. Interestingly, a low number of SNPs (<15 SNPs) was observed between environmental isolates retrieved ten years apart (Figure 12, in orange). Eight environmental isolates collected from different sink traps in the burn unit (ICU3) were closely related mostly with less than 10 SNP differences, except for environmental sample 13 retrieved in May 2013 which showed a slightly higher number of differences (<15 SNPs). Isolates belonging to the same patient, from Patient 2 and Patient 6, showed 4 and 5 SNP differences between them, respectively. The count of number of SNP differences observed between DLST 1-21 isolates depicted a low frequency of a highly variable number of SNPs, with a maximum of 114 SNPs (Figure 10).

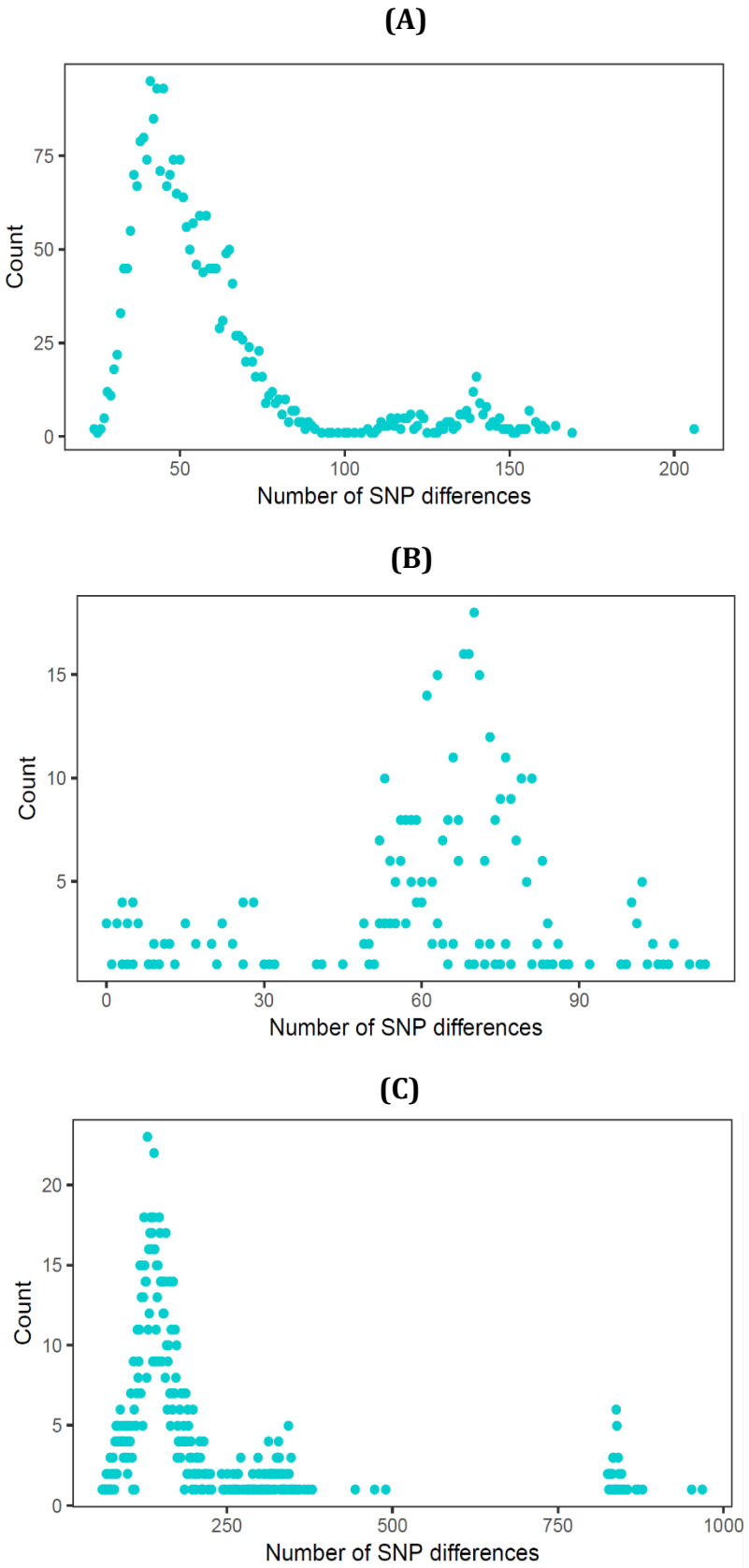


Figure 10. Frequency of number of SNP differences obtained with the standard methodology, mapping against *P. aeruginosa* PA14, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

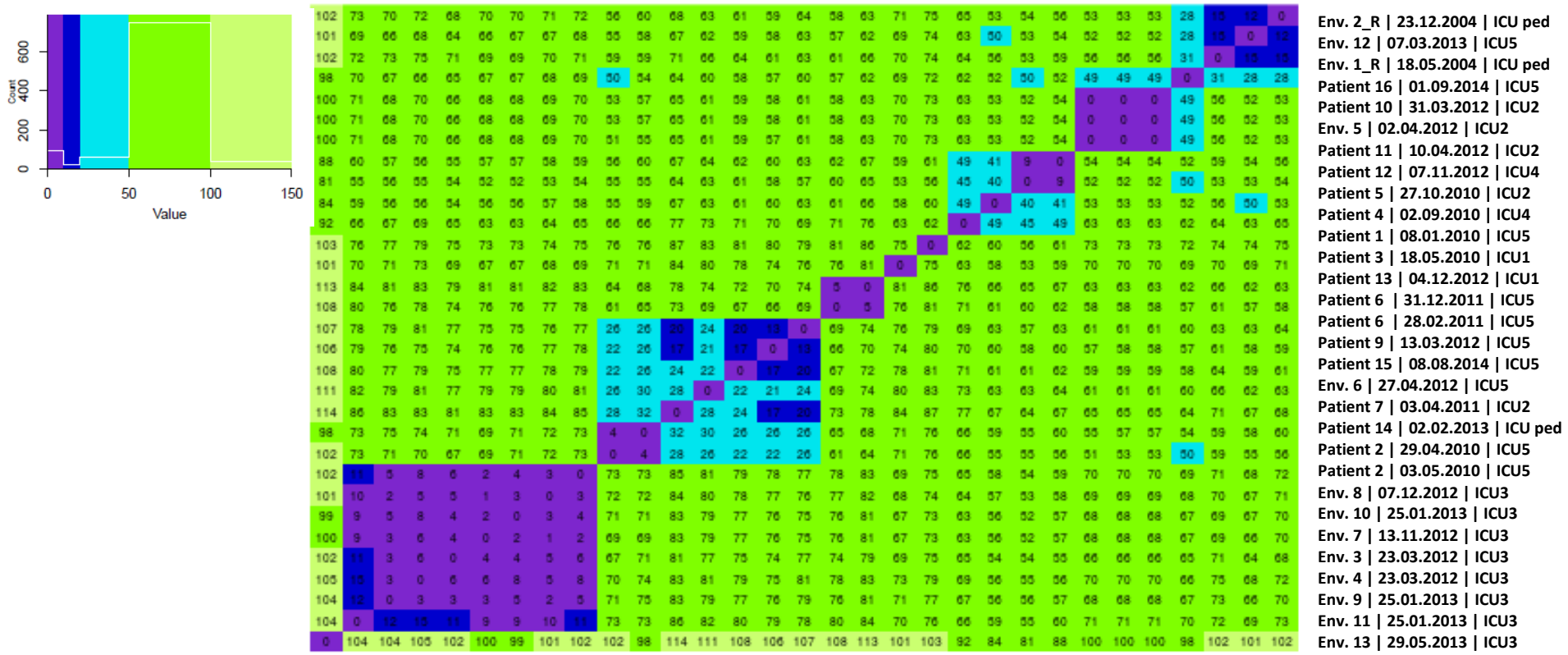


Figure 11. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the standard methodology, mapping against *P. aeruginosa* PA14. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 13 (first isolate) to Env.2_R (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

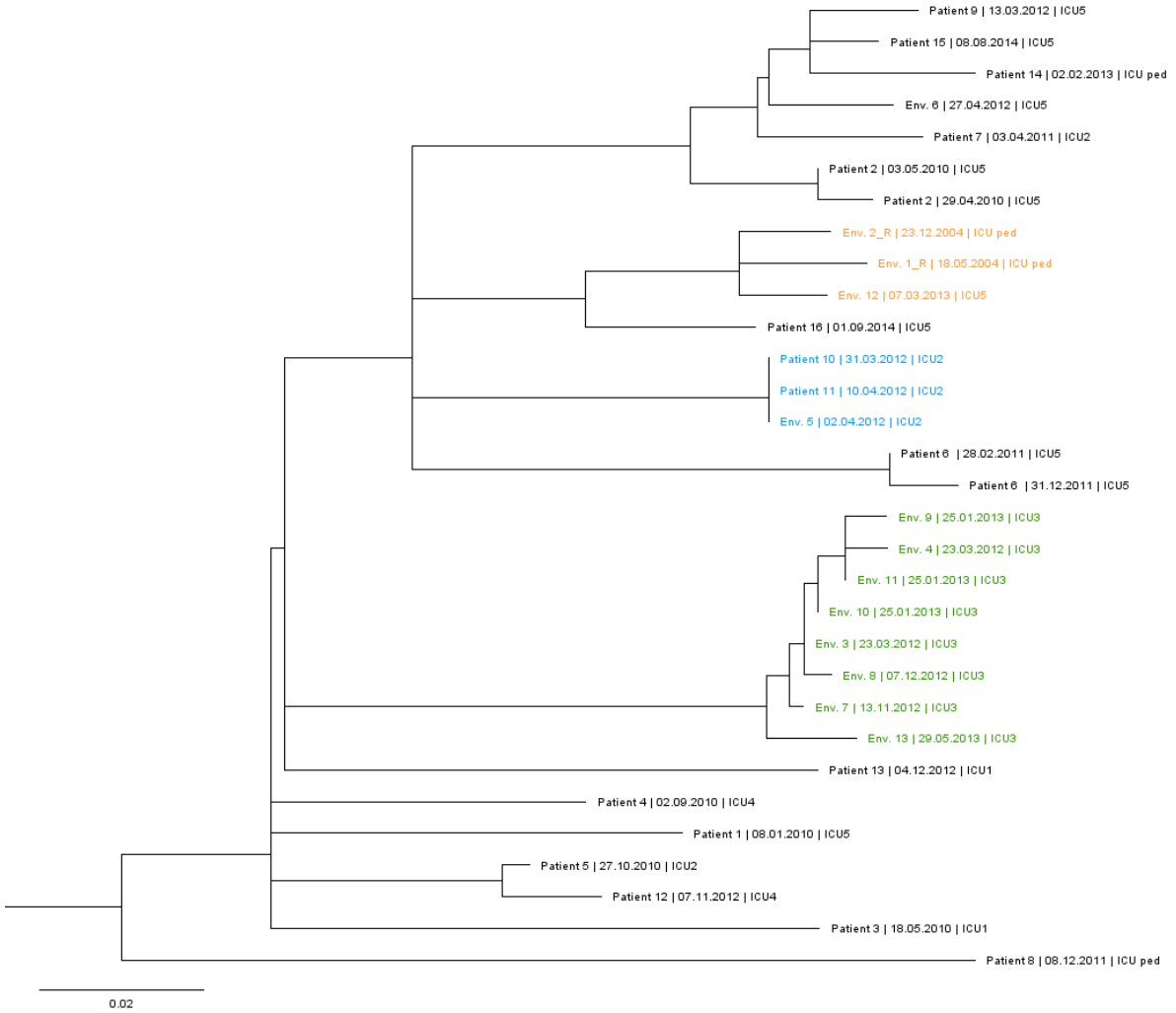


Figure 12. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the standard methodology, mapping against *P. aeruginosa* PA14. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 results showed that Patient 22 was distantly related to the remaining isolates with more than 800 SNP differences (Figure 13 and 14). One of the subclades represented in the phylogenetic tree (Figure 14) was composed in majority by isolates retrieved from the burn unit (Figure 14, in green). These isolates belonged to patients hospitalized during the same period in the burn unit. However, a high number of SNPs was still observed between them (>50 SNPs). Patient 11 and environmental isolates all retrieved from ICU2 clustered together in the phylogenetic tree (Figure 2.7, in blue), nonetheless with a high number of SNPs as observed before (>50 SNPs). One isolate from Patient 4 comprised in this subclade was associated with a long branch representative of more than 200 SNPs in relation to the remaining isolates. Another subclade harbouring five environmental isolates retrieved from both ICU3 and ICU4 clustered with the burn unit subclade with also more than 50 SNP differences. Isolates from the same patient retrieved less than two weeks apart had a high number of SNP differences between them, e.g. Patient 12 (85-124 SNPs). Two long branches respective to environmental samples 1 and 11 were detected (>200 SNPs). Figure 10 shows an elevated frequency of approximately 125 SNPs and in a smaller proportion from 750 to 1000 SNPs, suggesting most of the isolates were not closely related.

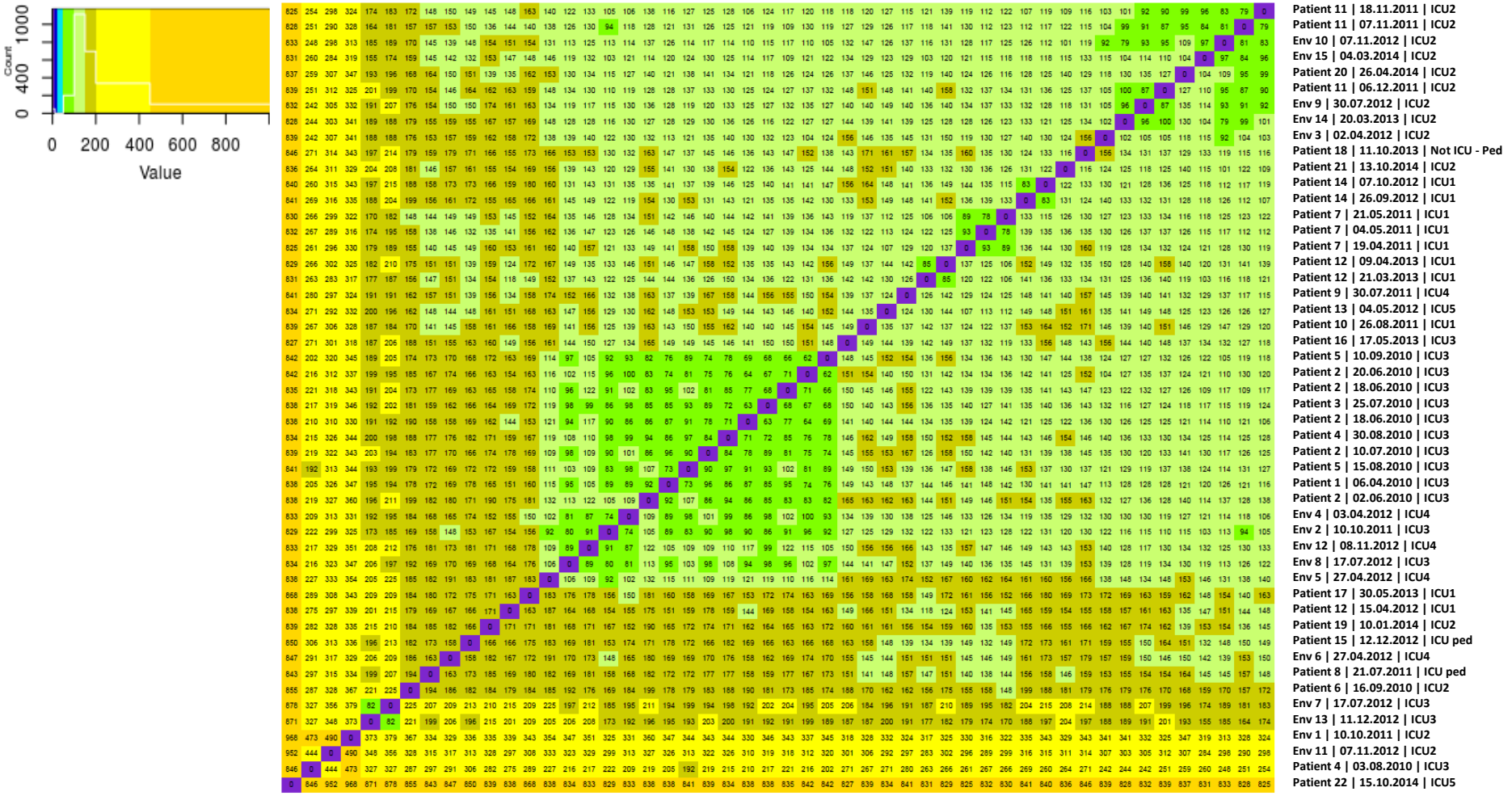


Figure 13. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the standard methodology, mapping against *P. aeruginosa* PA14. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 22 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, 400, and 1000. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

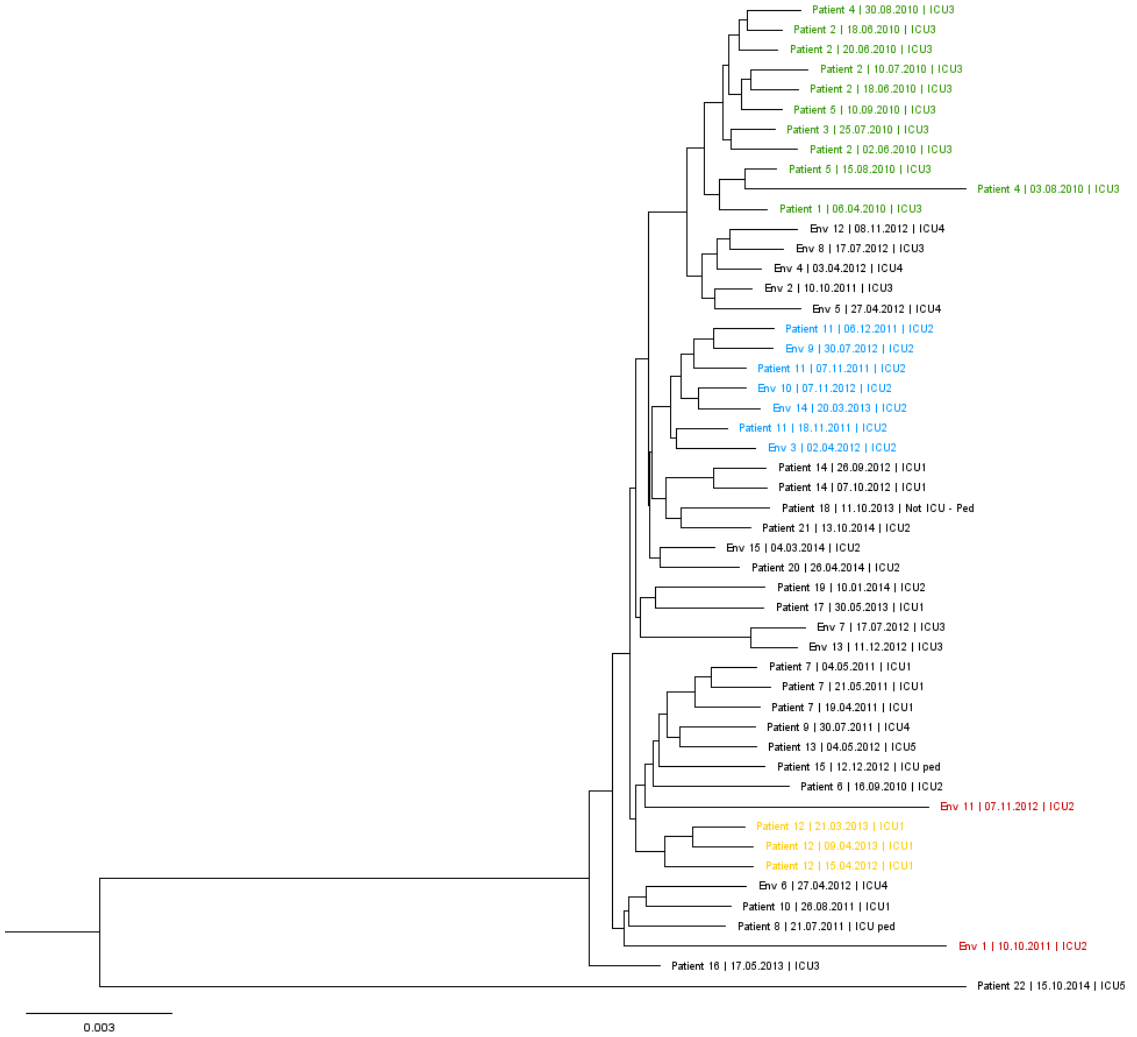


Figure 14. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the standard methodology, mapping against *P. aeruginosa* PA14. Patient 22 clustered apart from the remaining isolates. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red.

2.3.4. Adapted methodology with mapping against *P. aeruginosa* PA14

To understand if the odd number of SNPs observed in clusters 1-18 and 6-7 with the standard bioinformatic methodology was an accurate representation of our data or if it was due to artefacts, we decided to deconstruct the original methodology and complement it with additional steps. Only the isolates mentioned in the previous results for the first methodology will be used as example for a comparison measure between different methodologies. This information is summarized for each DLST type in Tables 2 to 4.

A drastic decrease in SNP differences was found for DLST 1-18 when analysed with the adapted methodology (Table 2).

Table 2. SNP differences (range) between DLST 1-18 isolates from the same patient when analysed with the standard and adapted methodology, by mapping against PA14 and PacBio references, when applying a mapping quality (MQ) value of 20 or 60, and when using a minimum of 20 or 10 reads to consider a SNP site.

Isolates	DLST 1-18									
	Standard methodology		Adapted methodology							
	PA14	PacBio	PA14				PacBio			
			20 MQ		60 MQ		20 MQ		60 MQ	
			20 reads	10 reads	20 reads	10 reads	20 reads	10 reads	20 reads	10 reads
Patient 1 (2 isolates)	48	12	14	14	11	11	13	13	12	12
Patient 4 (6 isolates)	31-73	0-6	2-9	3-9	0-6	0-7	0-8	0-8	0-7	0-7
Patient 23 (3 isolates)	81-121	2-5	8-11	10-13	2-5	2-5	2-5	2-5	0-2	0-2

Phylogeny of DLST 1-18 acquired with the adapted methodology demonstrated that isolates from Patient 1 were distantly related to the remaining collection, although with a slightly lower number of SNPs (91-101) (Figure 15 and 16). Several subclades were still observed (Figure 16), nonetheless most of the isolates shared only less than 10 SNPs except for one isolate belonging to Patient 16 (<13 SNPs). Five of six isolates from Patient 4 were present on the same subclade but all were closely related (<10 SNPs),

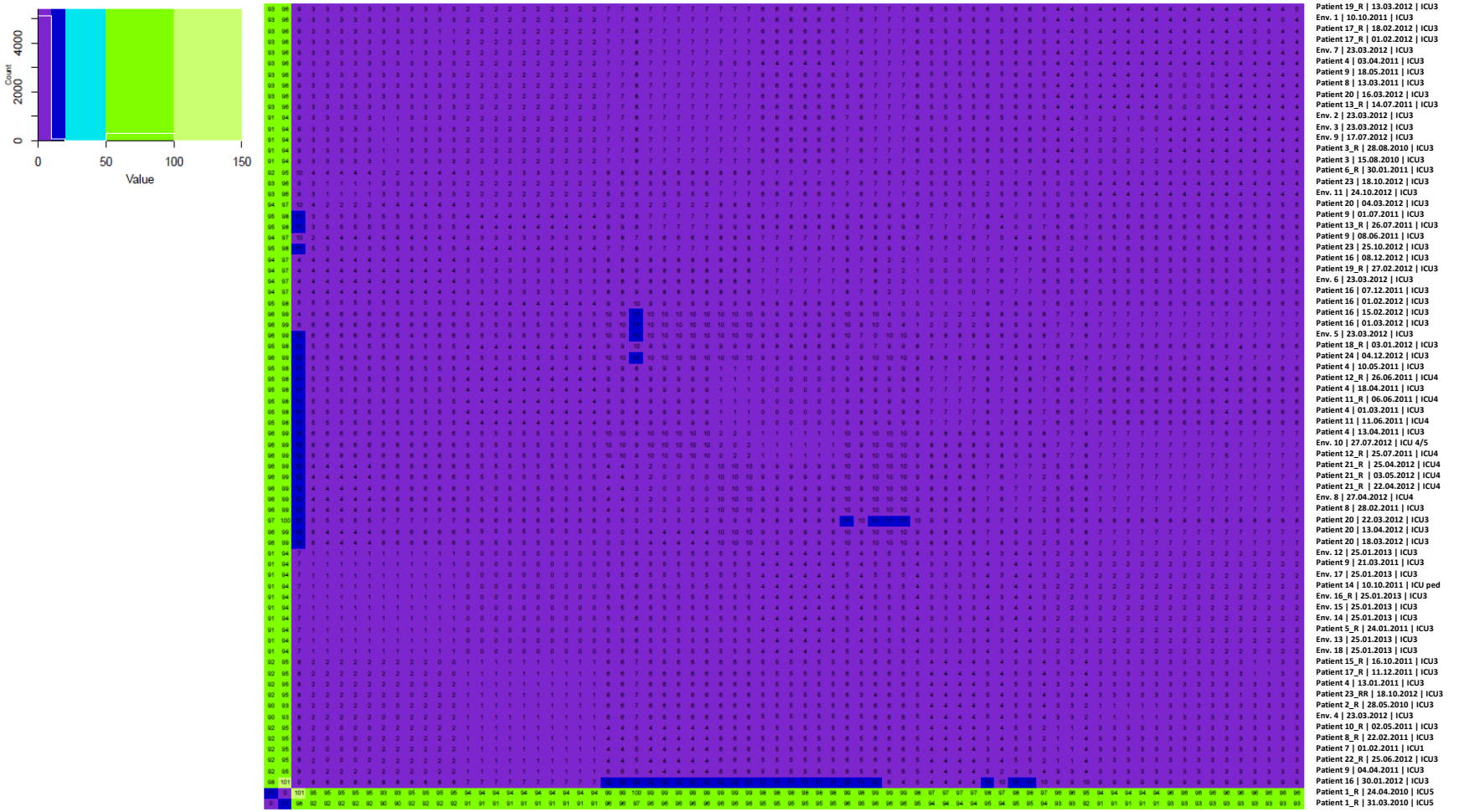


Figure 15. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Patient 19 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

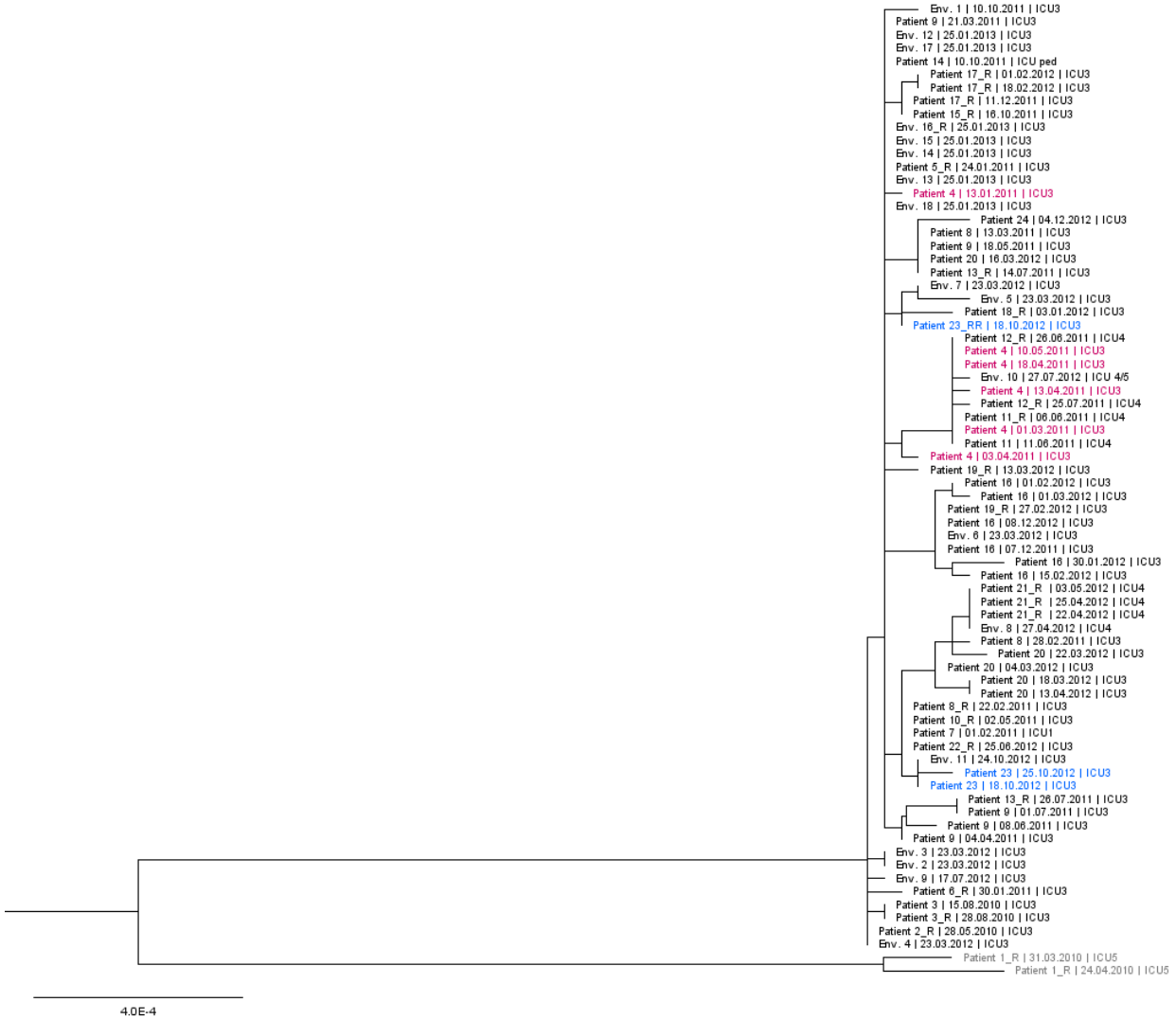


Figure 16. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey, and clustered apart from the remaining isolates. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

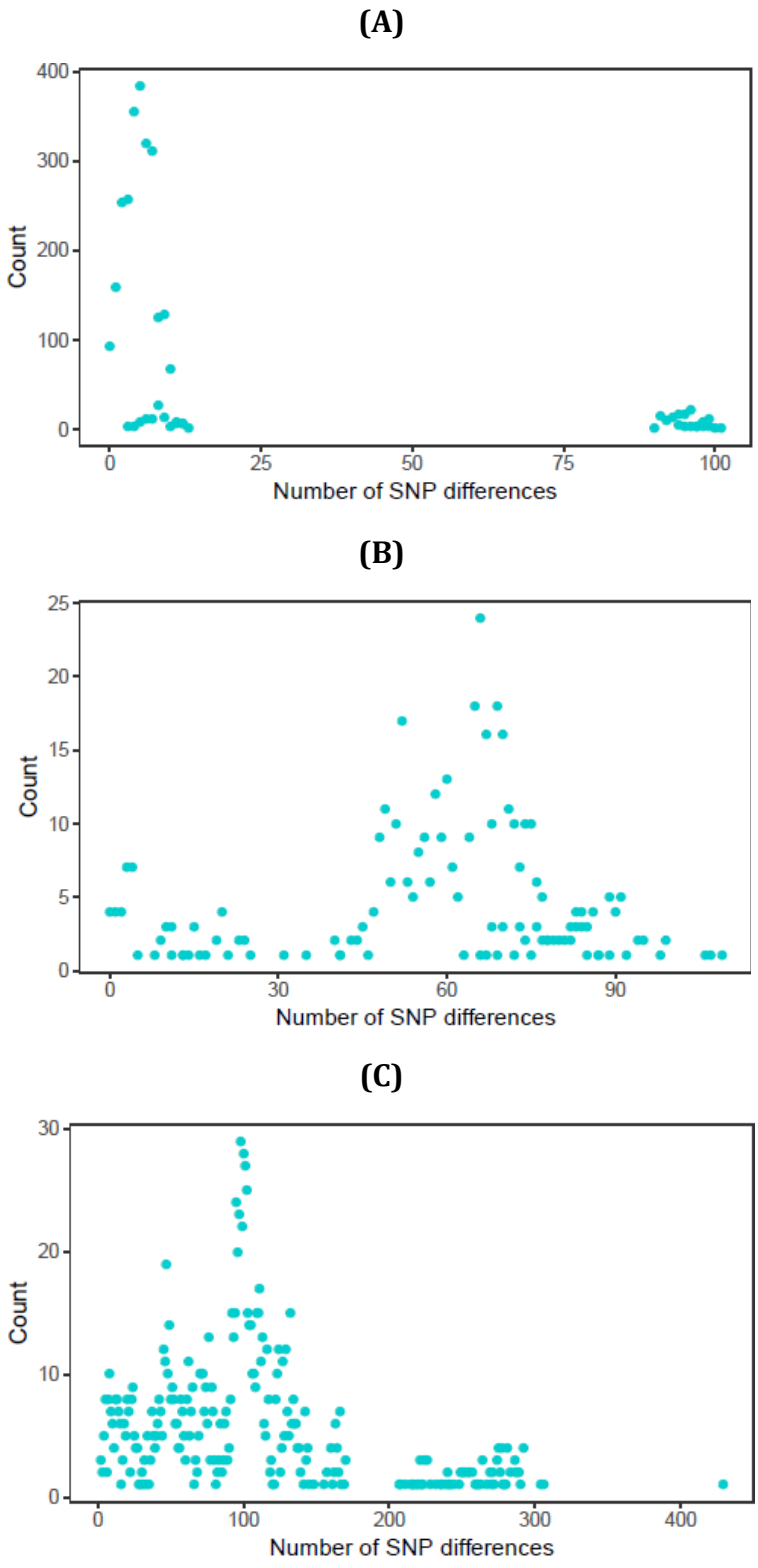


Figure 17. Frequency of number of SNP differences obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

Patient 23 isolates were separated in different subclades with 2 to five SNPs between them, and the long branch associated with a Patient 23 isolate (25 October 2012) was no longer observed. As clearly depicted on the graph of number of SNP differences frequency (Figure 17) minimum of 0 SNPs and a maximum of 101 SNPs (<13 SNPs excluding Patient 1) was found between DLST 1-18 isolates when analysed with the second pipeline.

Regarding DLST 1-21, no major differences were found between both methodologies (Figure 18 and 19). The small change in SNP differences obtained with the adapted methodology are listed in Table 3.

Table 3. SNP differences (range) between specific DLST 1-21 isolates when analysed with the standard and adapted methodology, by mapping against PA14 and PacBio references, when applying a mapping quality (MQ) value of 20 or 60, and when using a minimum of 20 or 10 reads to consider a SNP site.

Isolate	DLST 1-21									
	Standard methodology		Adapted methodology							
	PA14	PacBio	PA14				PacBio			
			20 MQ		60 MQ		20 MQ		60 MQ	
			20 reads	10 reads	20 reads	10 reads	20 reads	10 reads	20 reads	10 reads
ICU2 link (3 isolates)	0	1-3	0-1	0-1	0	0	0-1	0-1	0-1	1
Patient 5 and 12 (2 isolates)	9	6	5	5	4	4	6	6	6	6
Environmental isolates 10 years apart (3 isolates)	12-15	12-13	11-14	11-15	11-13	11-14	11-13	11-14	11-13	11-14
Subclade ICU3 (8 isolates)	<15	<12	<17	<17	<11	<11	<11	<11	<11	<11
Same patient	4-5	1-3	1-4	1-4	2-4	2-4	1-4	1-4	2-4	0-4

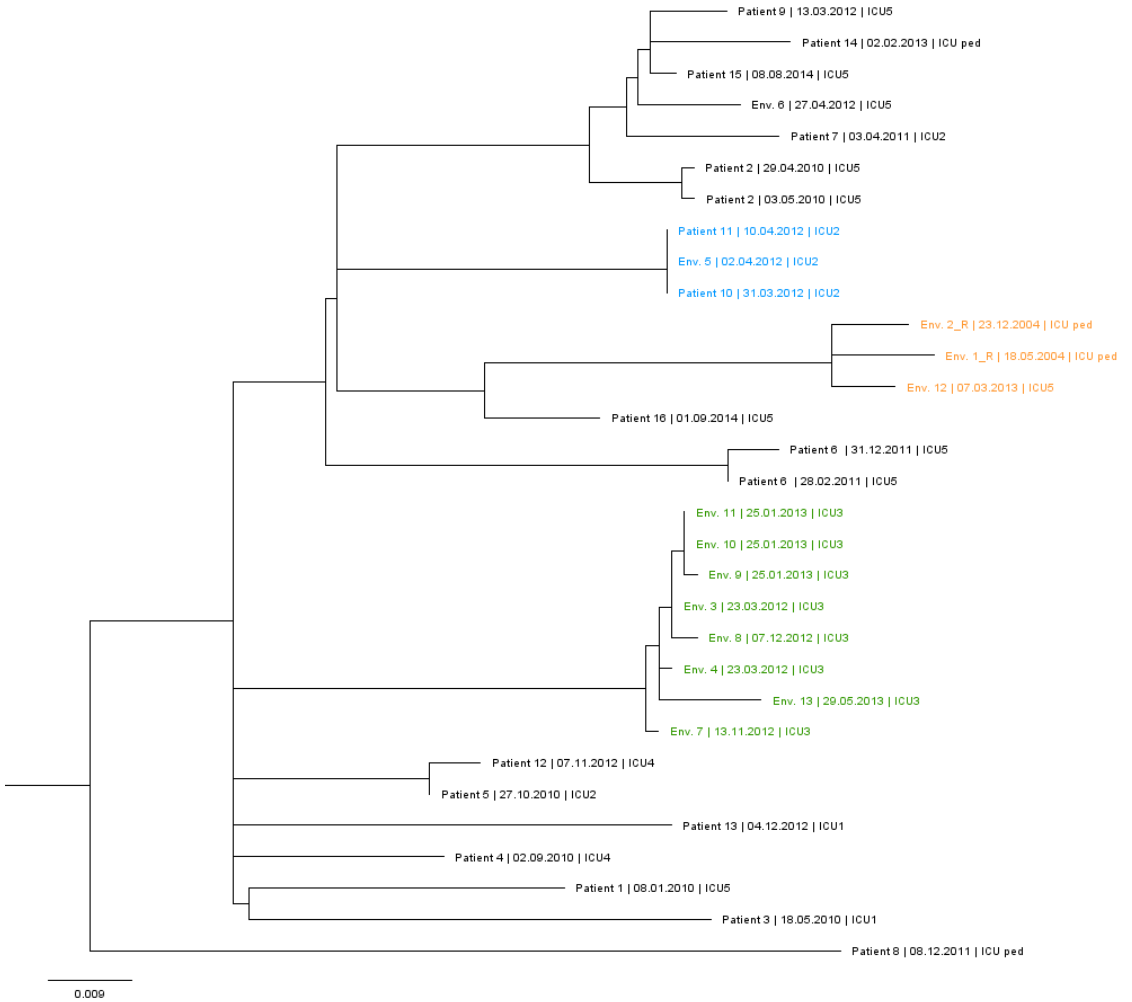


Figure 18. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

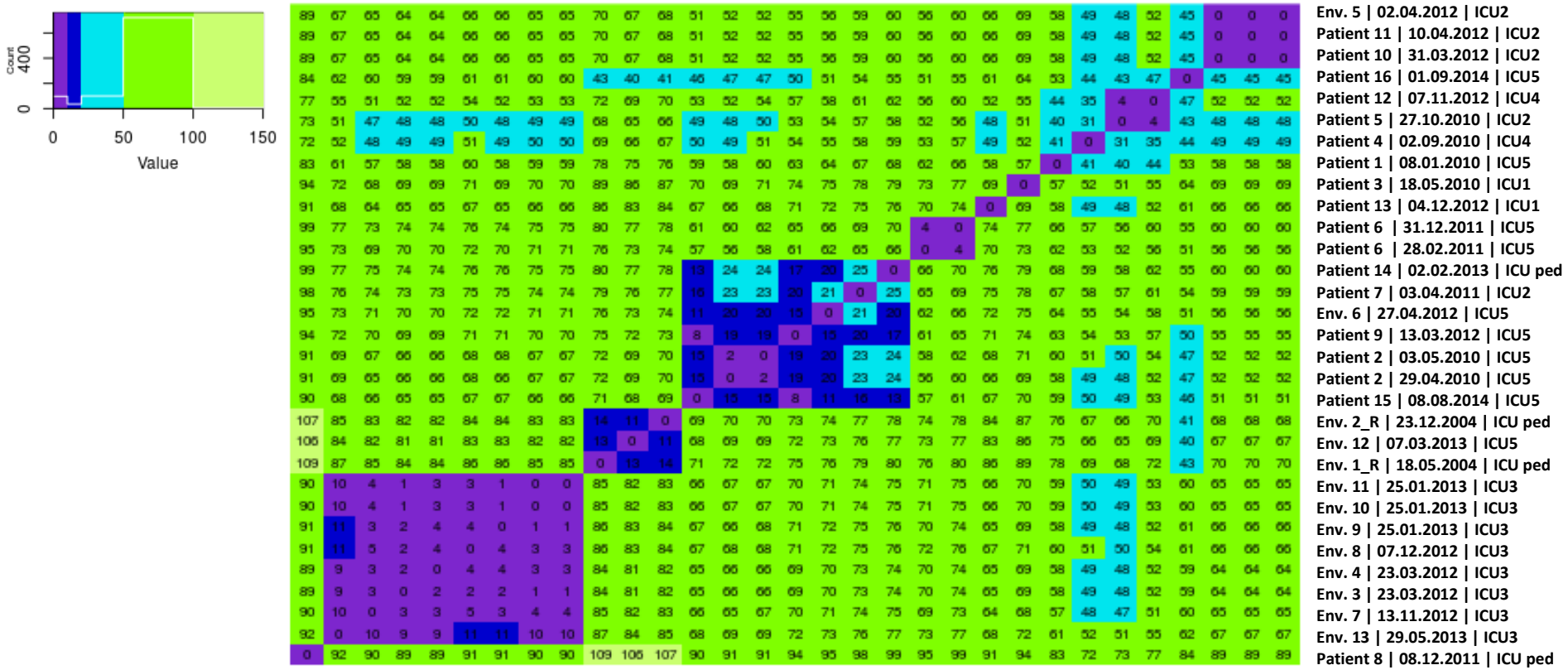


Figure 19. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Env. 5 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

Similarly to DLST 1-18, the number of SNPs between DLST 6-7 isolates also decreased greatly with the adapted methodology (Table 4). Patient 22 did not cluster apart from the remaining isolates although it still shared a high number of SNPs with them (70-279 SNPs) (Figure 20 and 21). Isolates retrieved from the burn unit suspected to be epidemiologically linked were genetically similar with less than 10 SNP differences. Environmental isolates retrieved from the burn unit and ICU2 which clustered with the burn unit subclade were 22 to 32 SNPs apart. A subclade composed of Patient 11 and environmental samples from ICU2 showed less than 14 SNPs as opposed to more than 50 SNPs acquired with the standard pipeline. The number of SNPs also decreased drastically between isolates from the same patient (7-30 SNPs). Patient 2 isolates (Figure 21, in orange) differed only by 7 to 12 SNPs. Figure 17 shows that most of DLST 6-7 patients were distantly related with 50 to 150 SNPs, with a minimum of 2 SNPs between isolates and a maximum of 429 SNPs.

Table 4. SNP differences (range) between specific DLST 6-7 isolates when analysed with the standard and adapted methodology, by mapping against PA14 and PacBio references, when applying a mapping quality (MQ) value of 20 or 60, and when using a minimum of 20 or 10 reads to consider a SNP site.

Isolate	DLST 6-7									
	Standard methodology		Adapted methodology							
	PA14	PacBio	PA14				PacBio			
			20 MQ		60 MQ		20 MQ		60 MQ	
			20 reads	10 reads	20 reads	10 reads	20 reads	10 reads	20 reads	10 reads
Burn unit subclade (11 isolates)	62-227	0-17	3-24	4-26	2-22	2-23	0-13	0-13	0-13	0-13
Subclade ICU2 (7 isolates)	81-135	2-14	1-12	1-12	3-14	3-14	0-9	0-9	0-7	0-7
Same patient	78-93	0-8	7-24	7-24	7-20	7-20	1-6	1-6	1-6	1-6
Patient 12 (3 isolates)	85-124	0-1	7-11	7-11	7-12	7-12	1-2	1-2	1-2	1-2

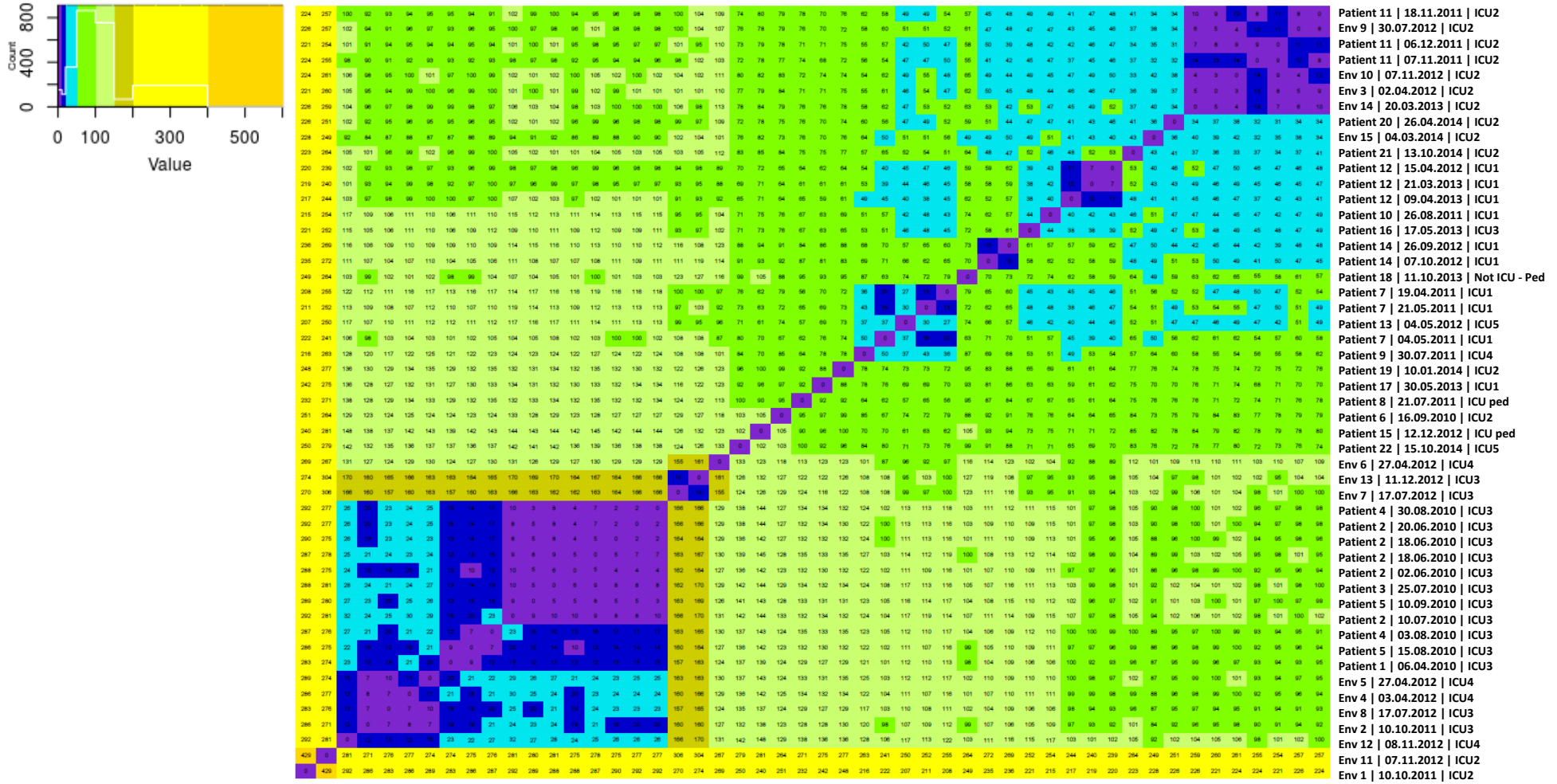


Figure 20. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100,150, 200, 400, and 600. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

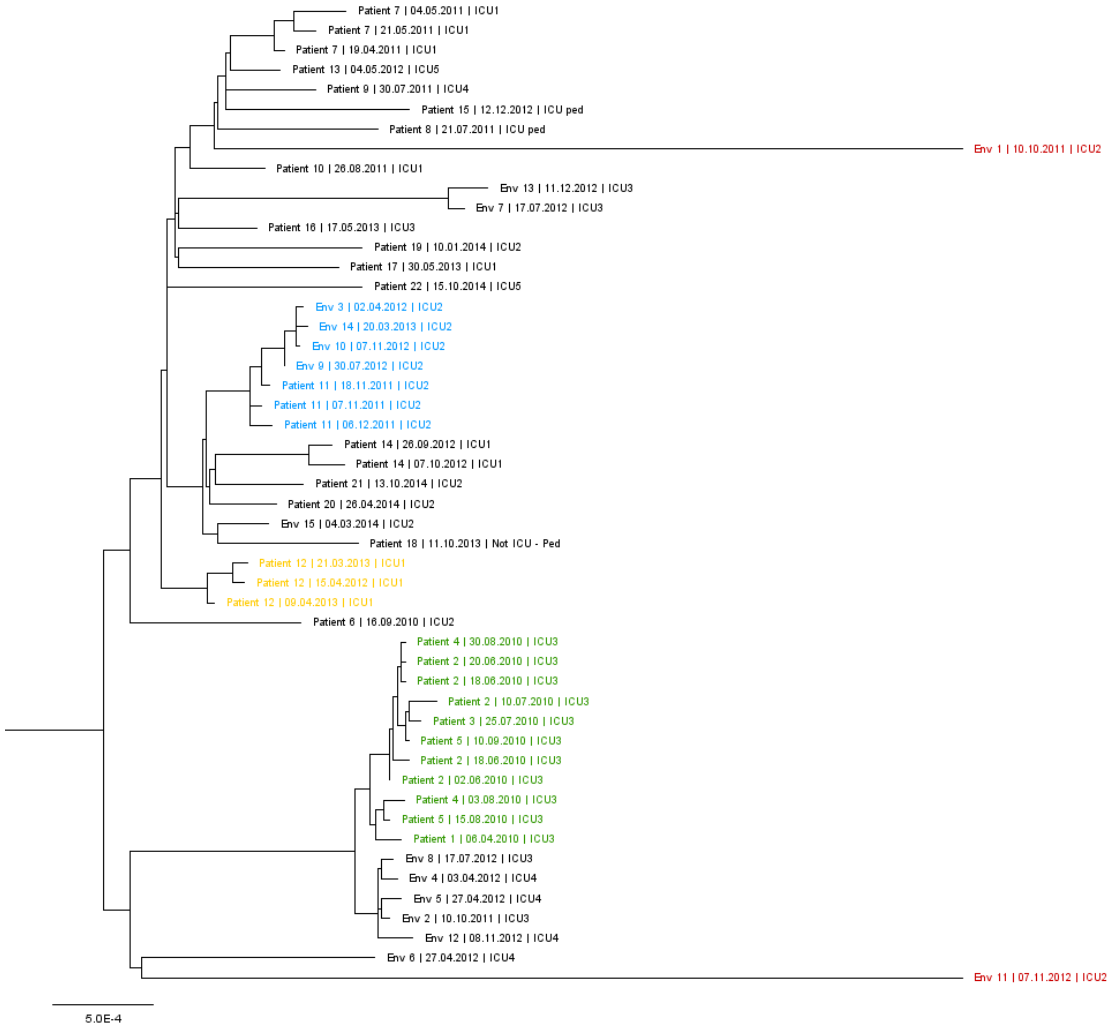


Figure 21. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 10 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red.

2.3.5. Standard methodology with mapping against the PacBio reference

At the start of this study, only available complete genomes were published belonging to ST253 (DLST 1-21), and not for the other two ST found in our isolate collection: ST1076 (DLST 1-18) and ST17 (DLST 6-7). Thus, we decided to use the well-known reference genome from *P. aeruginosa* PA14 (ST253) for the mapping step in both tested bioinformatic pipelines. However, it is known that the use of a closely related reference genome to the isolate collection can overcame challenges during the analysis and help to determine with more accuracy outbreak and non-outbreak isolates (2). Considering the weight of the reference in the analysis we decided to construct a reference genome from the index case of each DLST cluster, as described in Section 3.6. Here we present the results of WGS analysis with the standard methodology using the corrected PacBio references instead of *P. aeruginosa* PA14, also summarized in Tables 2 to 4.

Interestingly, the use of a closely related reference genome decreased greatly the number of SNPs between DLST 1-18 isolates, as it was observed when using the stricter adapted methodology (Table 2). Isolates from Patient 1 were considered again genetically distant from the remaining isolates (105-118 SNPs) (Figure 22). Although the tree topology was similar to the ones observed in previous results (Figure 22), the number of SNPs between most of the isolates was less than ten, with only a minority having less than 15 SNP differences between them. Once again, five out of six isolates from Patient 4 clustered together with less then 13 SNP differences. Patient 23 isolates were separated on the phylogenetic tree although only by 2 to 5 SNPs, and no long branch associated with one of the isolates was observed. The minimum number of SNPs between DLST 1-18 isolates was zero and maximum was 118 (Figure 24), with the rest of the isolates being genetically similar with less then 13 SNPs apart.

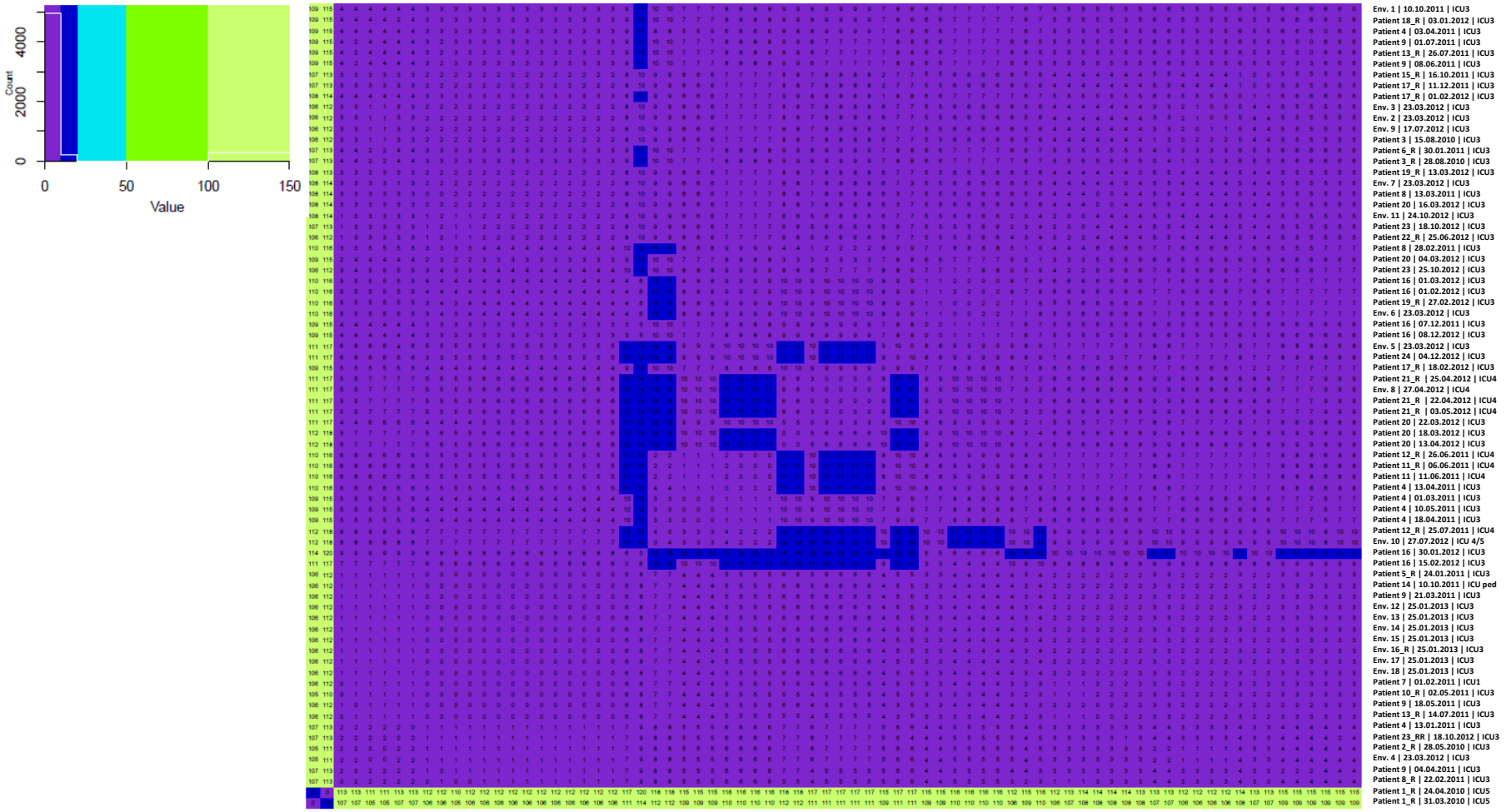


Figure 22. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the standard methodology, mapping against the PaCBio reference. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Env.1 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

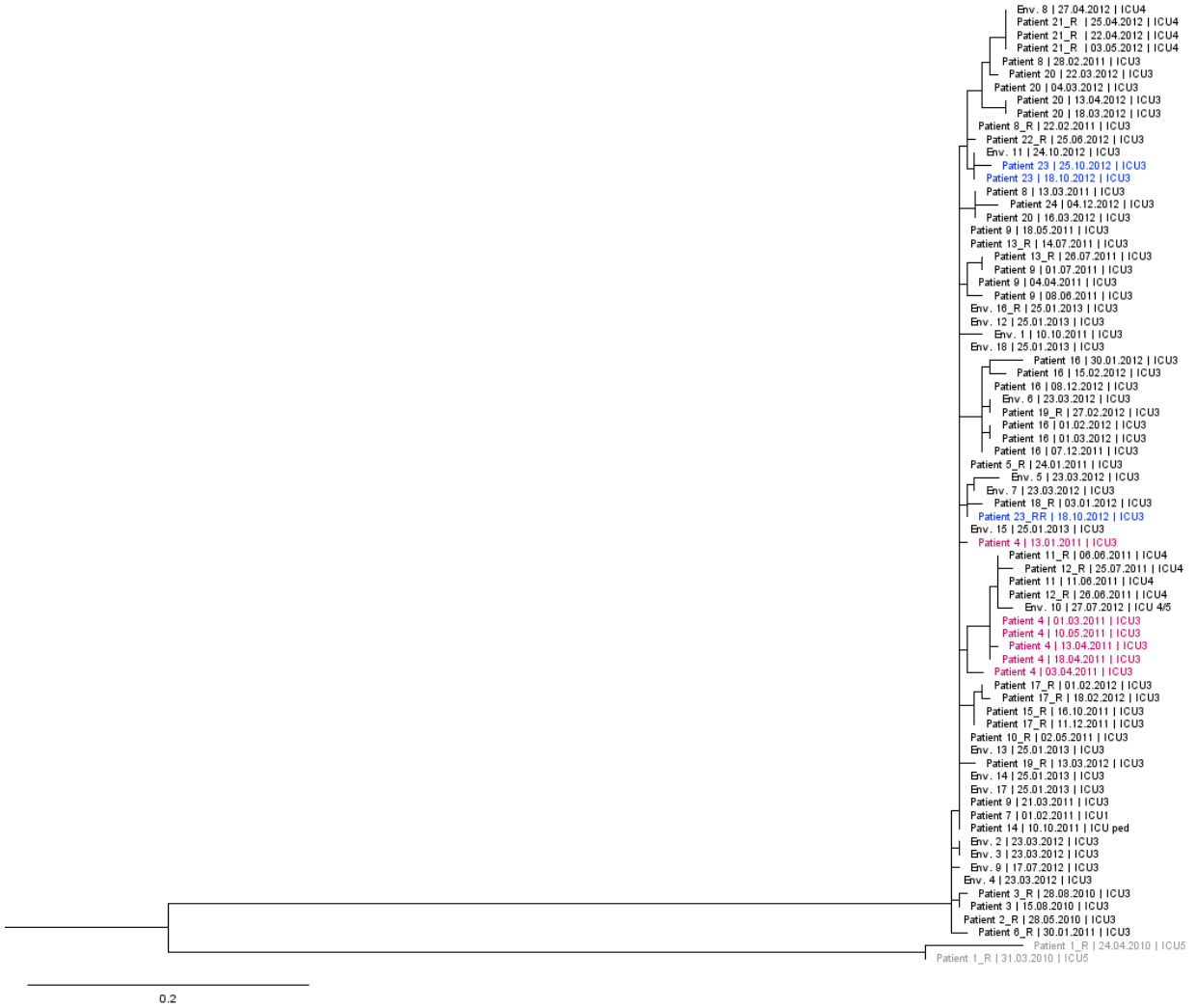


Figure 23. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the standard methodology, mapping against the PacBio reference. Non-outbreak isolates belonging to Patient 1 are highlighted in grey, and clustered apart from the remaining isolates. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

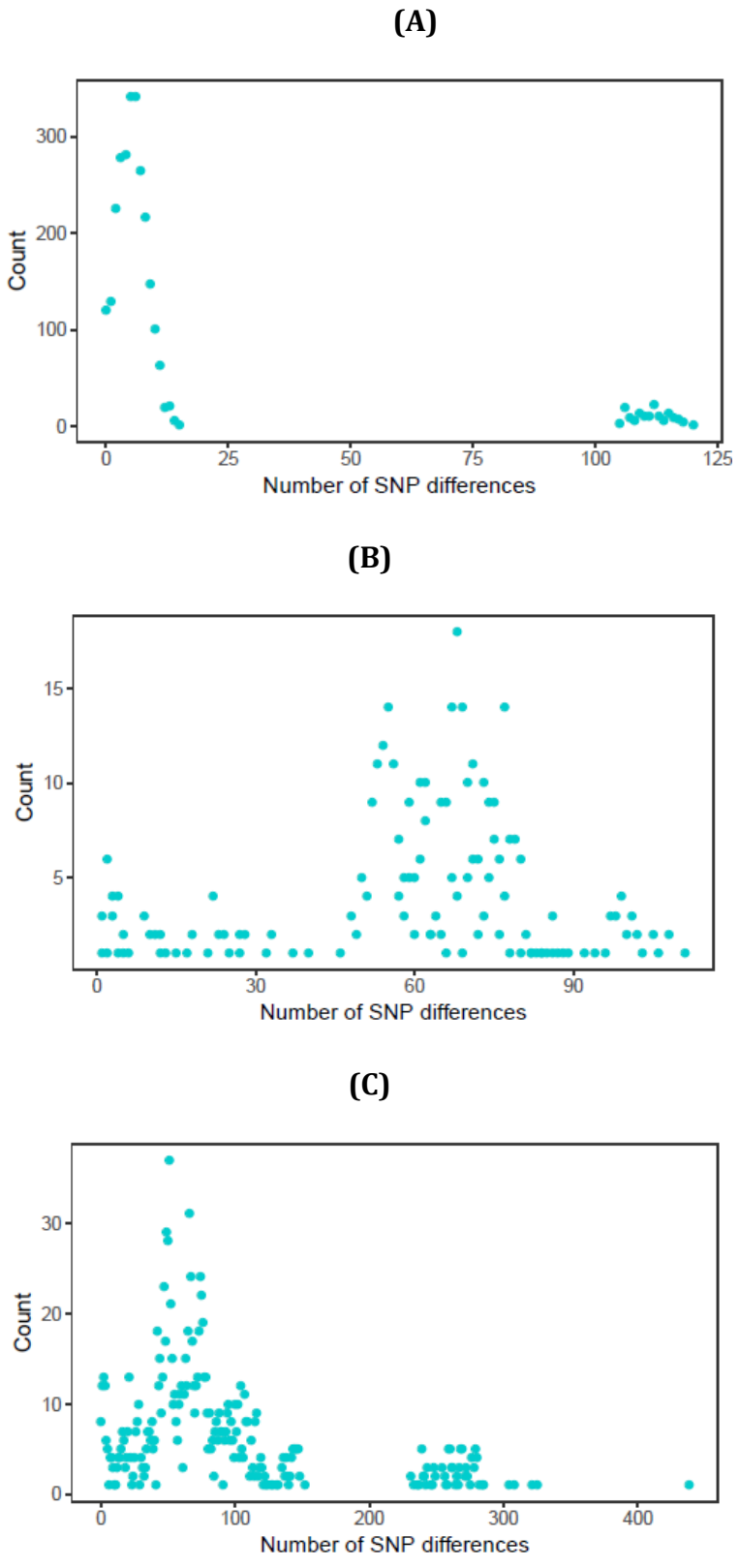


Figure 24. Frequency of number of SNP differences obtained with the standard methodology, mapping against the PacBio reference, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

For DLST 1-21, applying the original methodology with mapping against the PacBio reference gave similar results to when both methodologies were performed with *P. aeruginosa* PA14 (Figure 25 and 26). However, one difference was in the number of SNPs between isolates suspected to be epidemiologically linked which instead of zero, showed 1-3 SNPs differences. Additional minor changes in SNPs are present in Table 3.

This methodology also lead to a lower number of SNPs between DLST 6-7 isolates (Figure 27 and 28). Results were very similar to the ones achieved with the adapted methodology using PA14, except for environmental sample 1 being genetically distant from the remaining isolates and not considered just a long branch as before. Other slight SNPs differences in the results are present in Table 4, as no additional clear change was detected.

2.3.6. Adapted methodology with mapping against the PacBio reference

As was done for the previous bioinformatic schemes, the SNP differences of specific isolates discussed in the results are summarized in Table 2 to 4. Patient 1clustered far apart from the remaining isolates with a maximum of 120 SNPs between a pair of isolates (Figure 29 and 30). Several subclades were identified comprising both clinical and environmental isolates, however most of the outbreak isolates were closely related with less than 10 SNP differences. Isolates retrieved from the same patient, although some were separated on the phylogenetic tree, were genetically different by only less then 10 SNPs differences, e.g. Patient 4 (<7 SNPs) and Patient 23 (0-2 SNPs). The number of SNP differences count (Figure 31) demonstrated that the maximum of differences observed between a pair of isolates is 16 and the minimum is zero, except for isolates belonging to Patient 1 (120 SNPs).

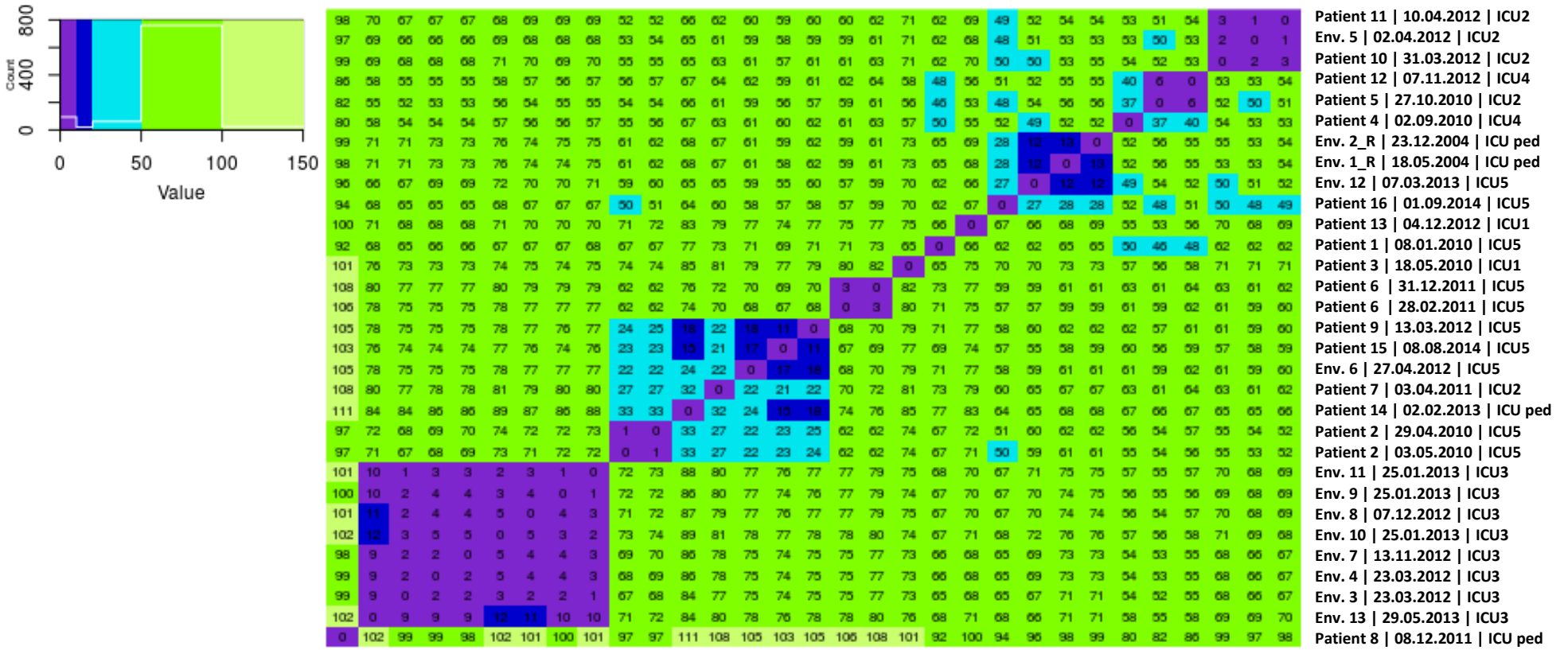


Figure 25. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the standard methodology, mapping against the PaCBio reference. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

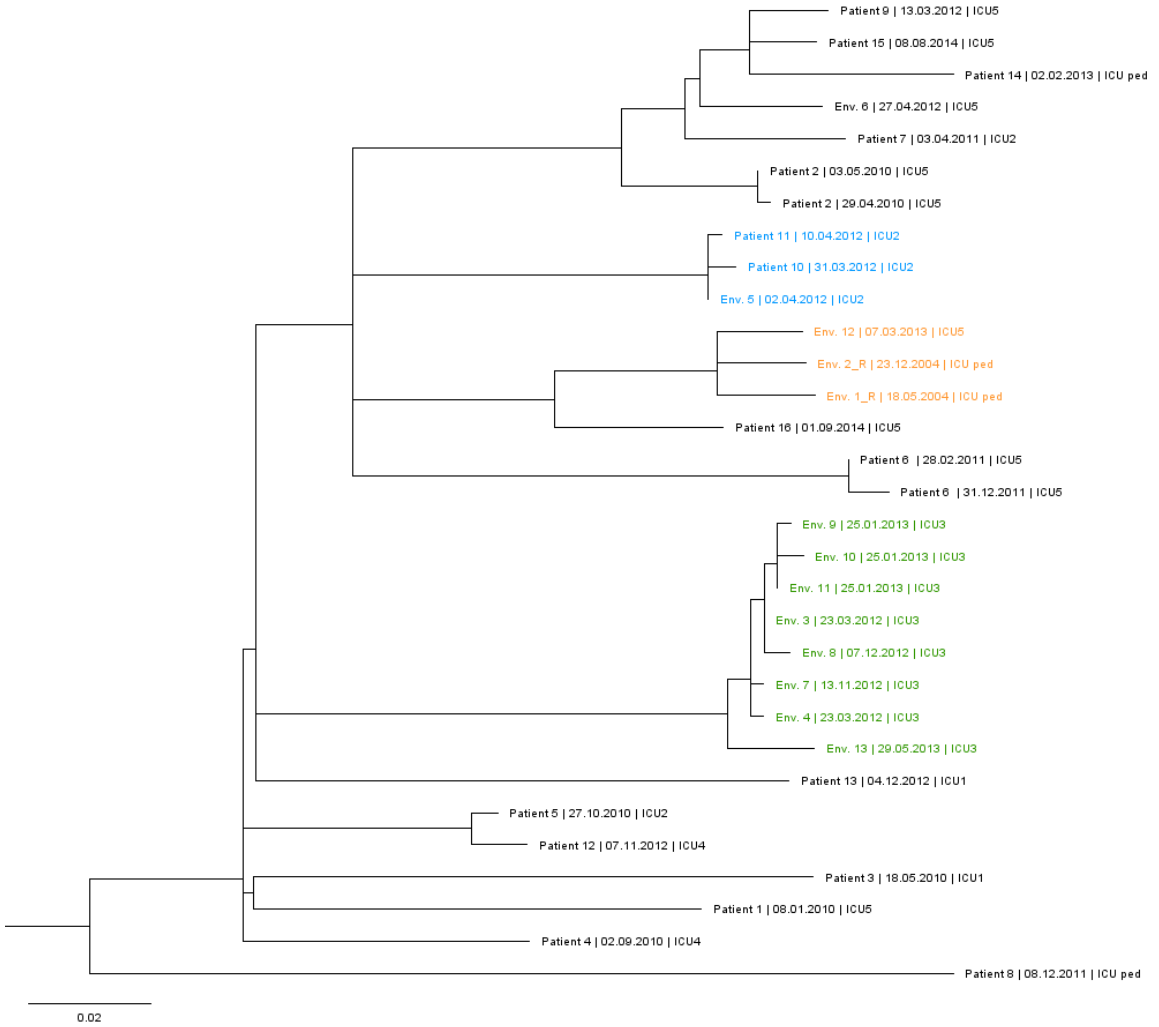


Figure 26. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the standard methodology, mapping against the PacBio reference. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

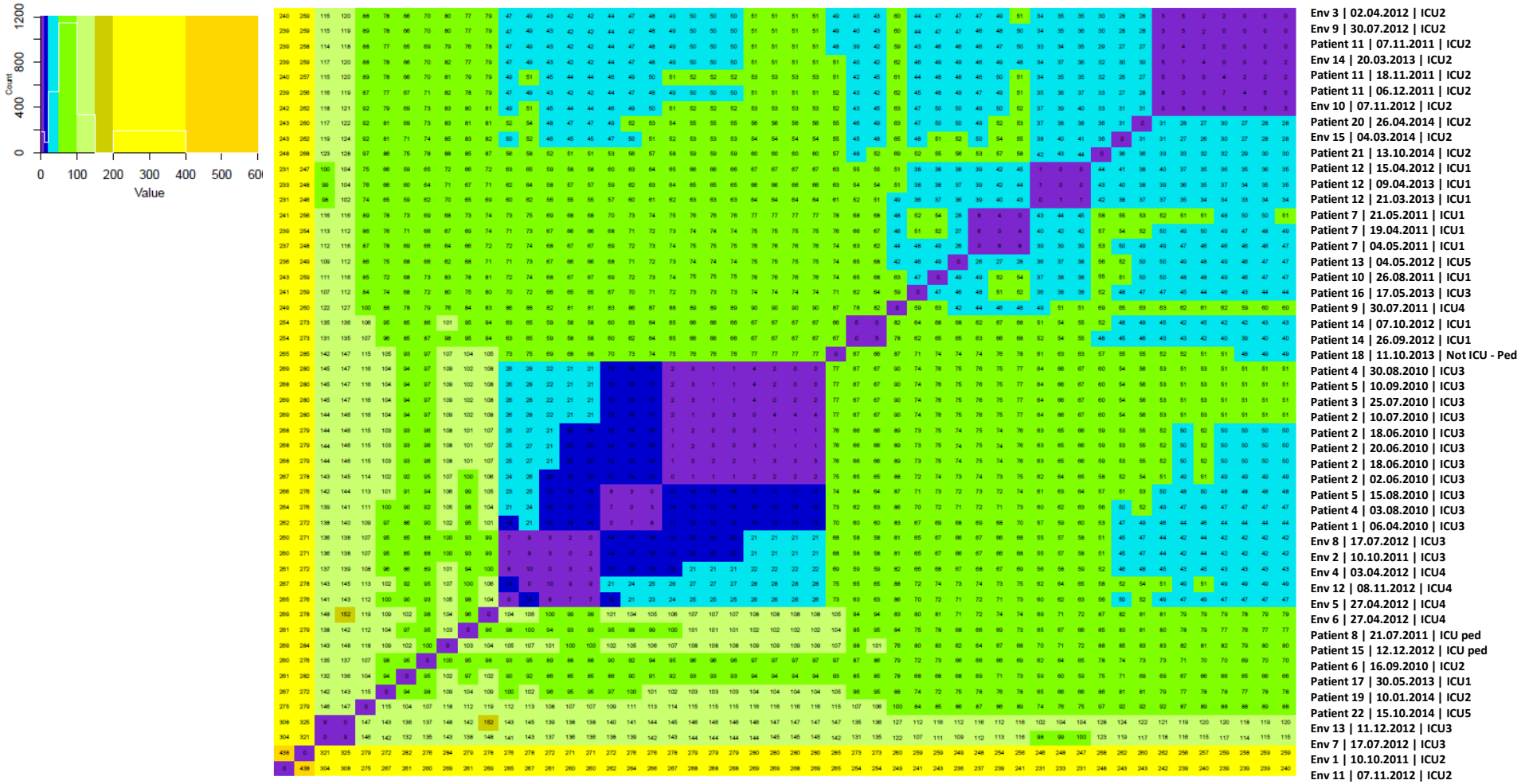


Figure 27. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the standard methodology, mapping against the PaCBio reference. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env 11 (first isolate) to Env 3 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150,200, 400, and 600. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

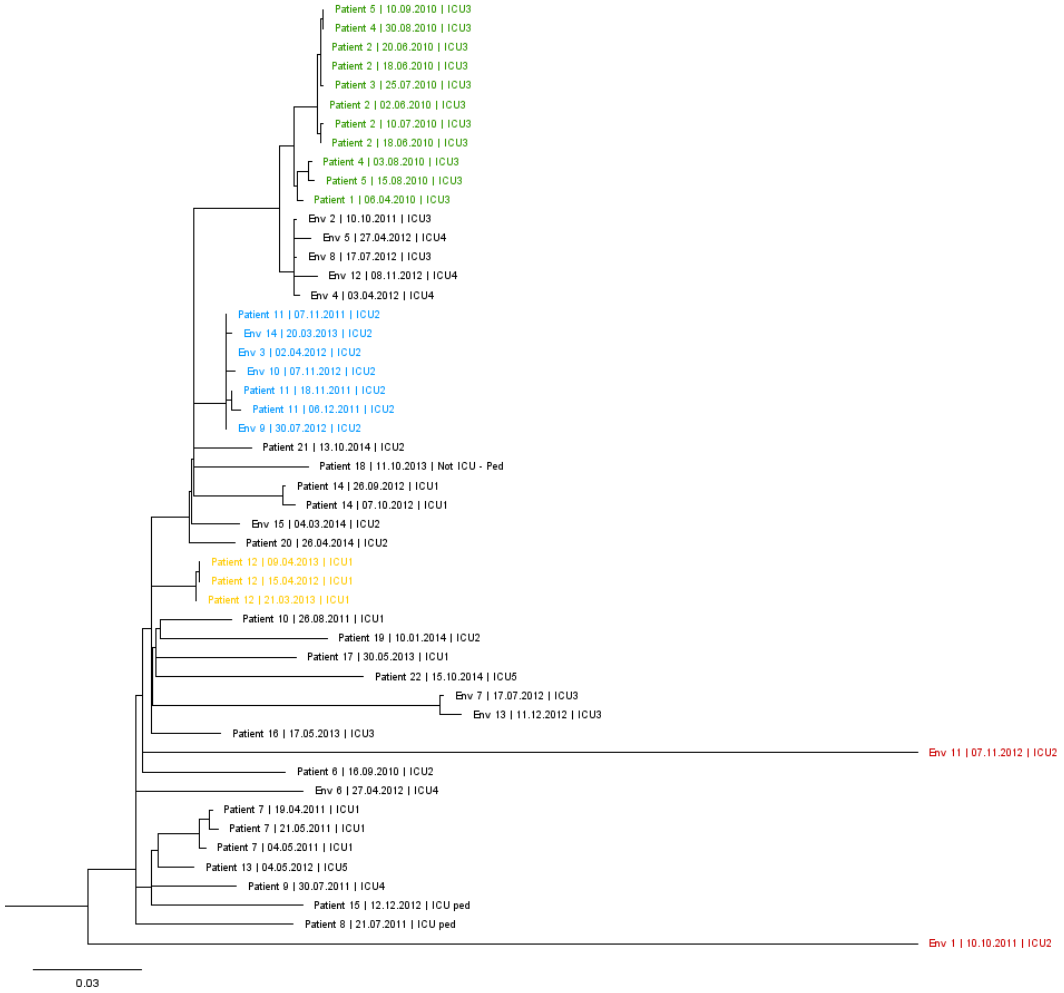


Figure 28. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the standard methodology, mapping against the PacBio reference. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red.

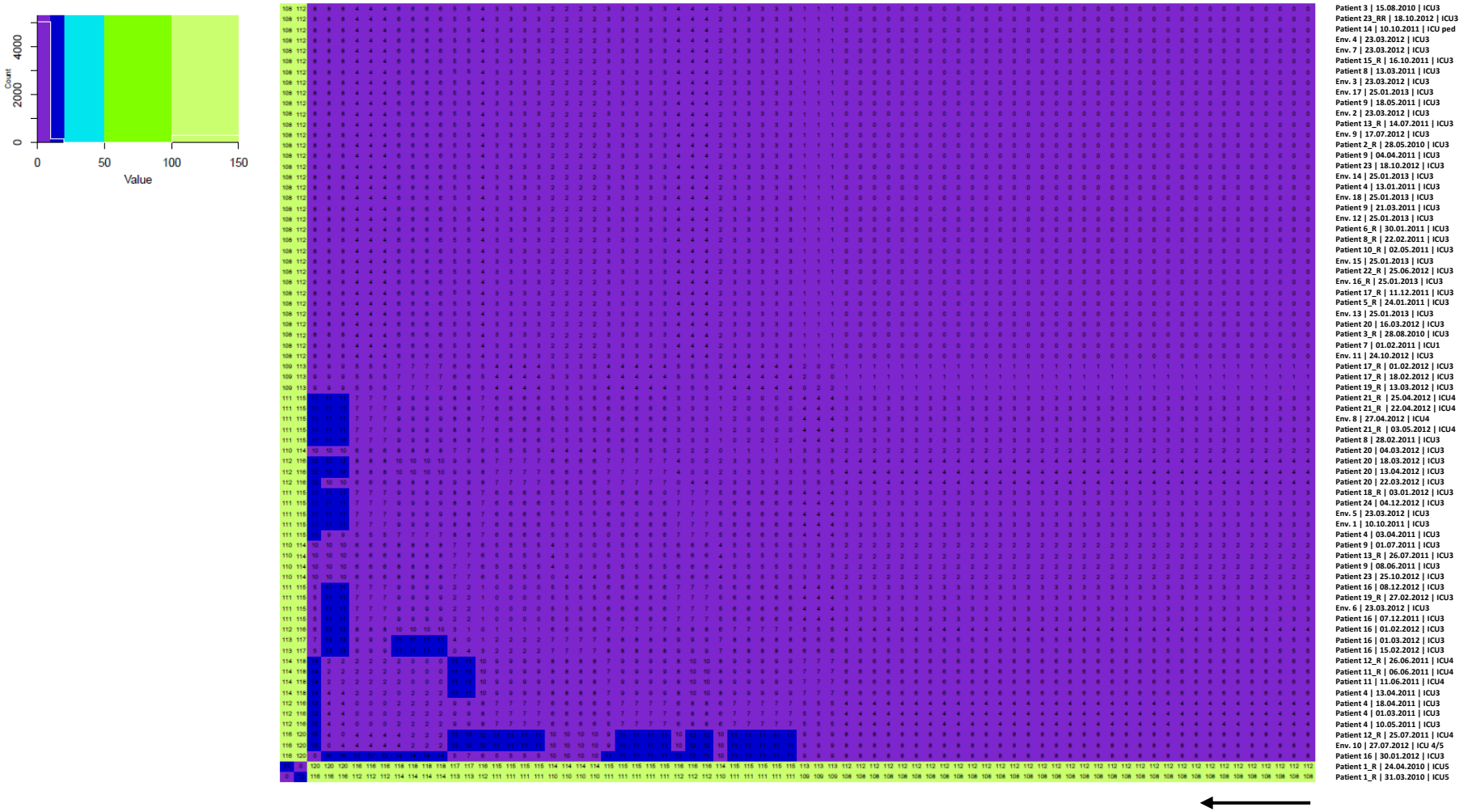


Figure 29. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PaCBio reference with a mapping quality of 60 and with a minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Patient 3 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

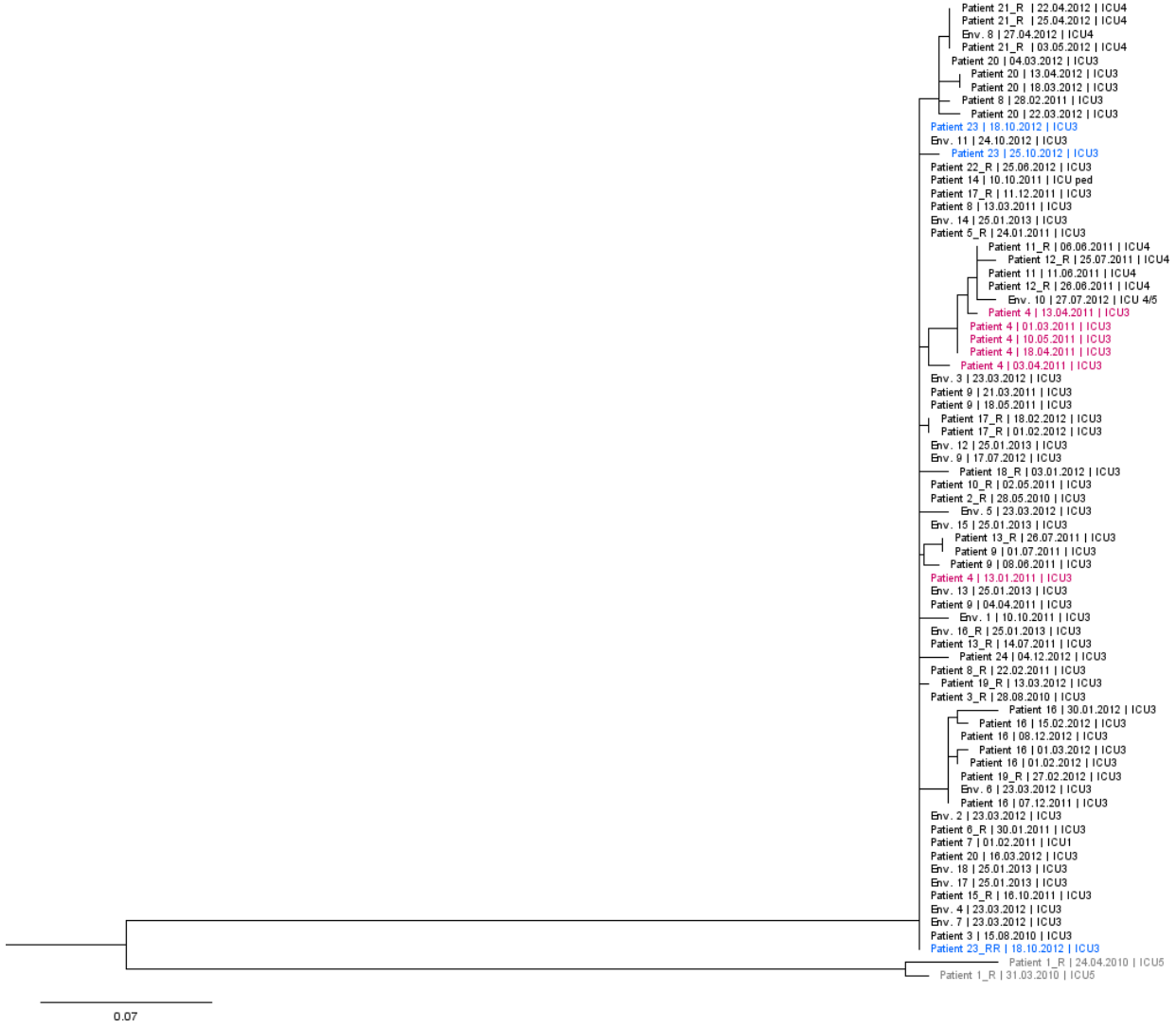
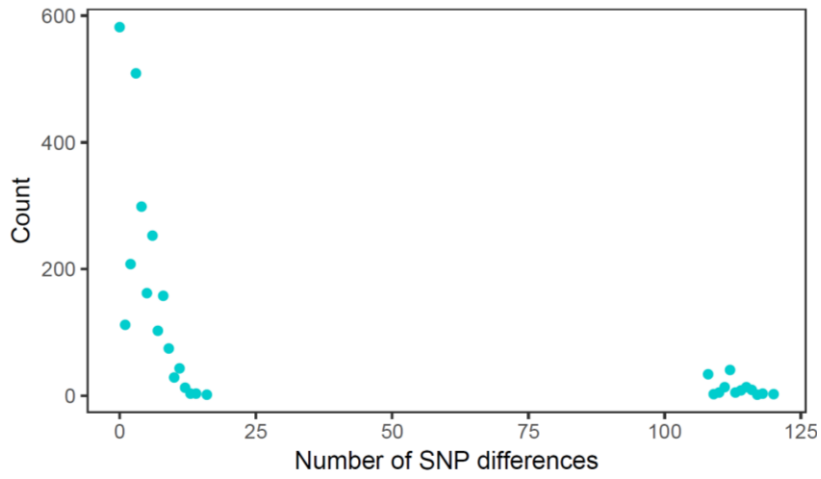
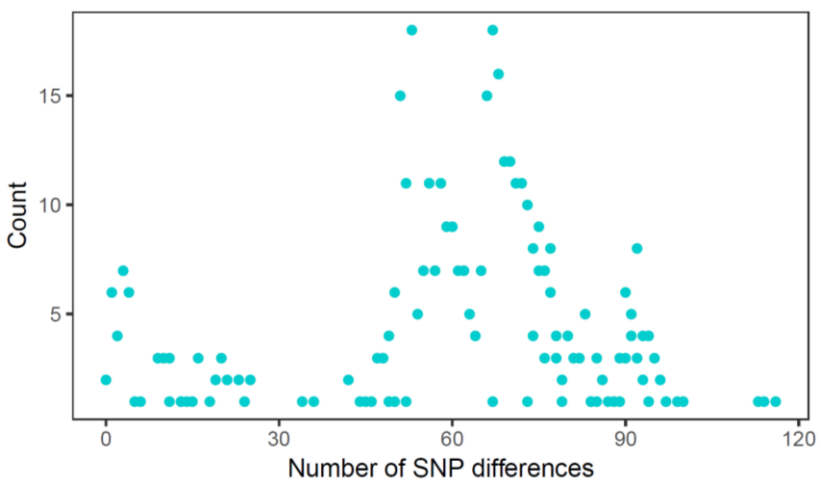


Figure 30. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacBio reference. Non-outbreak isolates belonging to Patient 1 are highlighted in grey, and clustered apart from the remaining isolates. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

(A)



(B)



(C)

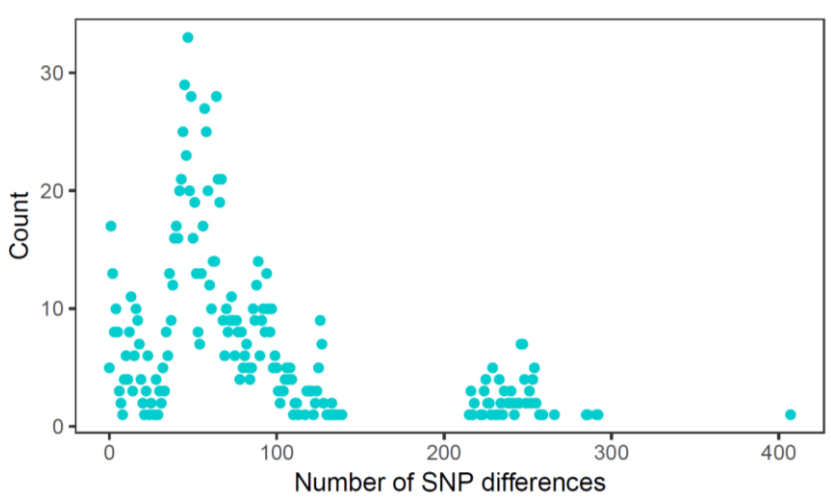


Figure 31. Frequency of number of SNP differences obtained with the adapted methodology, mapping against the PacBio reference, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

For DLST 1-21, two isolates from two patients (zero SNPs) and one isolate from the environment (Figure 32 and Figure 33, in blue,) showed one SNP difference between them. Only six SNP differences were found between two isolates of two patients hospitalized in different ICUs (ICU2 and ICU4), retrieved two years apart. A low number of SNPs (<14 SNPs) was observed between environmental isolates retrieved ten years apart (highlighted in orange). Environmental isolates retrieved from different sink traps in the burn unit (ICU3) clustered together with less than 11 SNP differences. Isolates belonging to the same patient were closely related with a maximum of 4 SNPs between them. SNPs count for this DLST cluster demonstrated a high variability of the number of SNPs found, ranging from zero to 116 SNPs between a pair of isolates (Figure 31).

Figure 35 representing DLST 6-7 cluster phylogeny showed several clades and subclades with high number of SNP differences (Figure 34). One subclade (Figure 35, in green) was composed by isolates retrieved in the burn unit with less than 13 SNPs differences. These isolates belonged to patients for which epidemiological links were suspected. Four environmental isolates retrieved from both the burn unit and ICU4 were closely related to the burn unit cluster (10-19 SNPS). A adapted subclade was constituted by closely related isolates from Patient 11 sampled in ICU2 and environmental isolates from the same ICU (0-7 SNPs). All isolates recovered from the same patient clustered together with only a few SNPs differences (1-6), e.g. Patient 12 (1-2 SNPs) (Figure 35, in yellow). Long branches with more than 200 SNPs associated with two isolates from the environment (Env 1, 11) were detected. Most isolates had 50 to 150 SNP differences between them, with a maximum of 429 SNPs and a minimum of 2 SNPs (Figure 31).

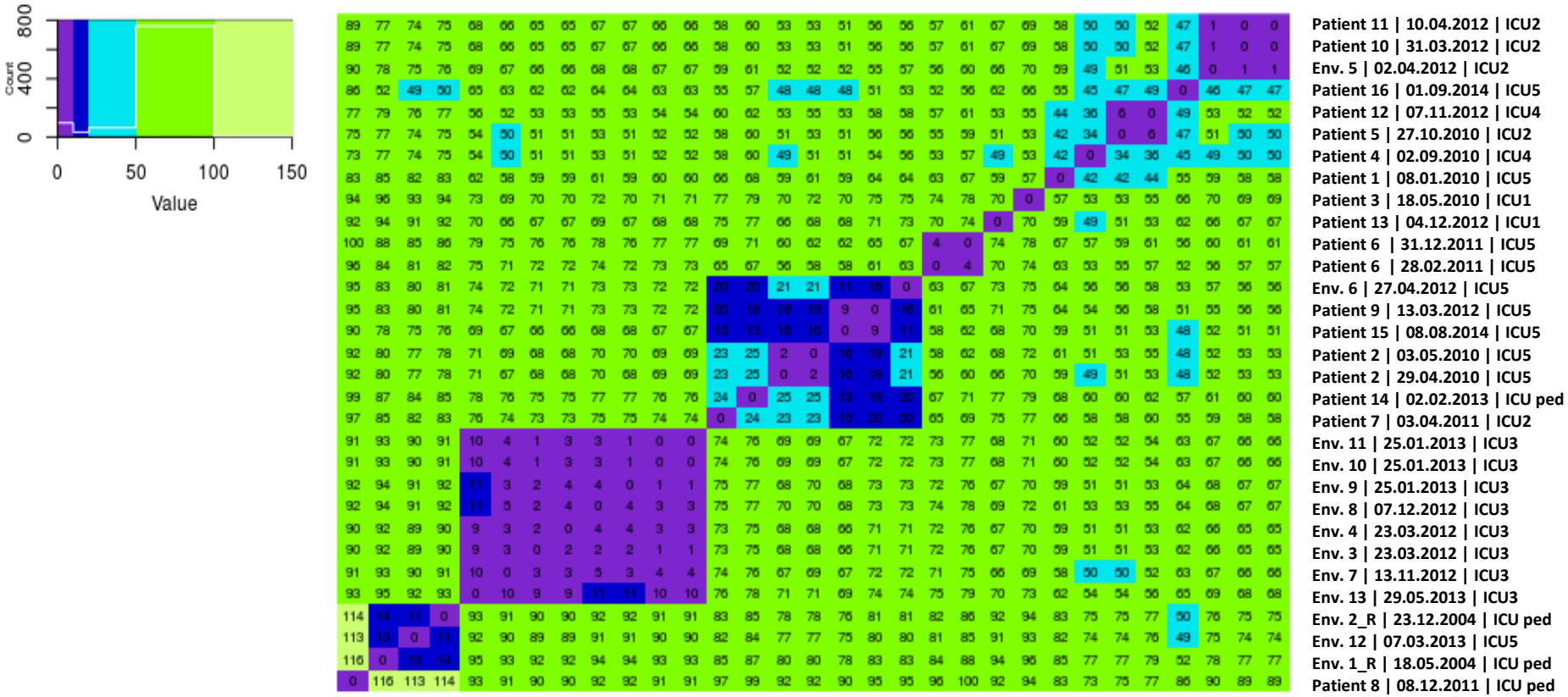


Figure 32. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PaCBio reference with a mapping quality of 60 and with a minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

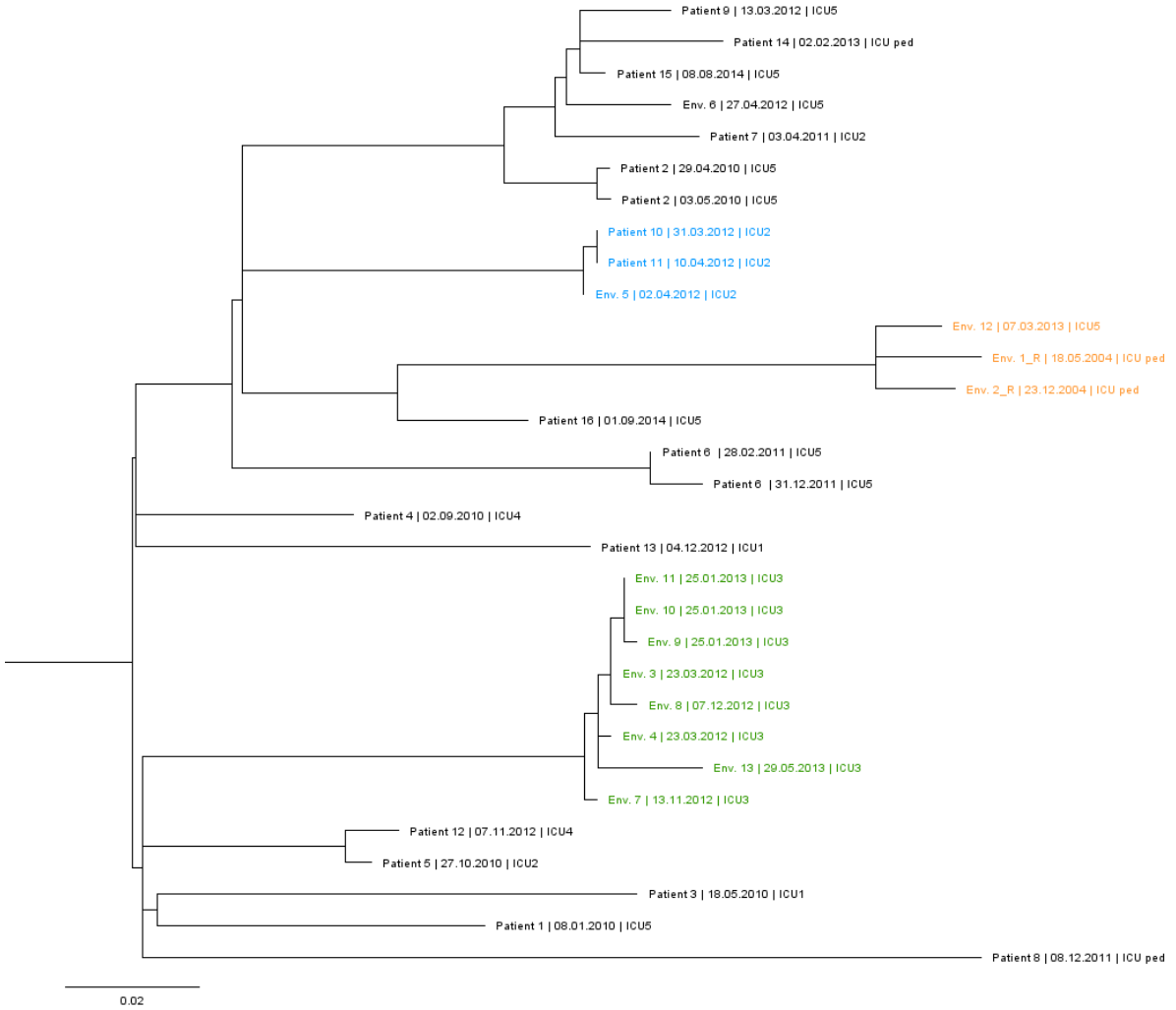


Figure 33. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PaCBio reference with a mapping quality of 60 and with a minimum of 10 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange;

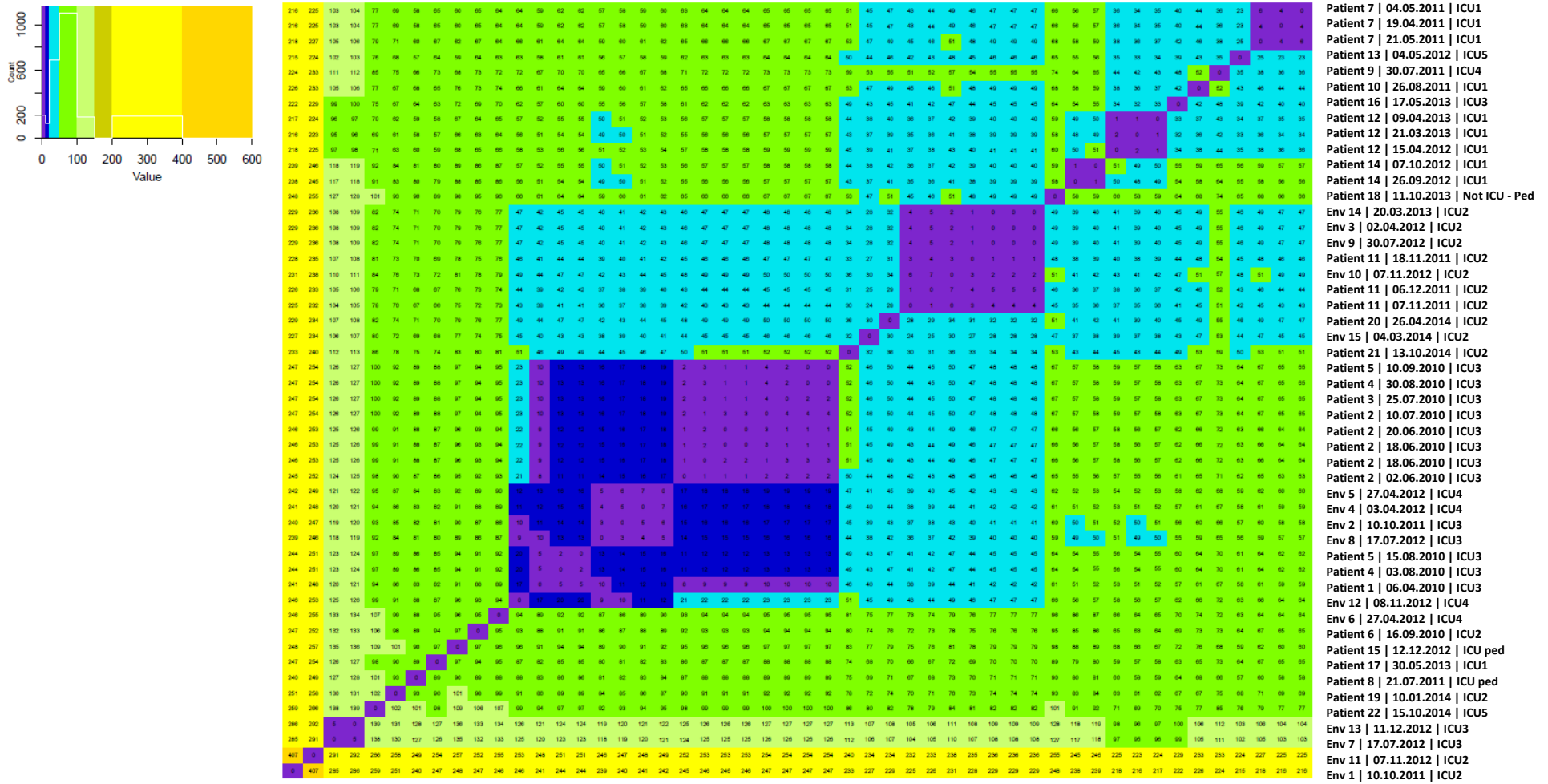


Figure 34. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PaCBio reference with a mapping quality of 60 and with a minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Patient 7 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, 400, and 600. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

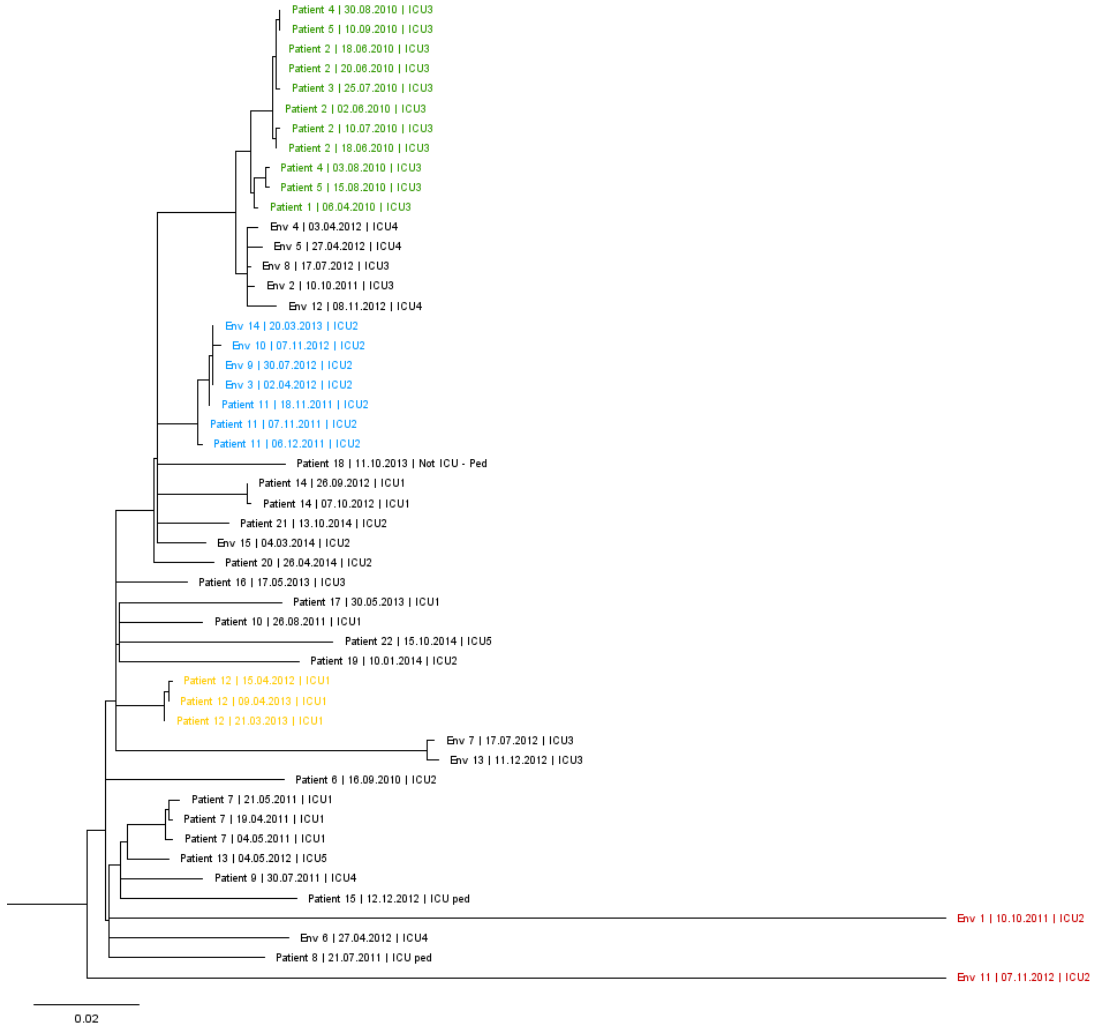


Figure 35. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PaCBio reference with a mapping quality of 60 and with a minimum of 10 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red.

2.3.7. Adapted methodology: different parameters

2.3.7.1. Variant calling using a lower mapping quality threshold

Minimum mapping quality is an important parameter used to exclude low-confidence alignments. It is a probabilistic value that determines the confidence of the read to be correctly mapped to the genomic coordinates of the reference genome. When developing the second bioinformatic scheme, thresholds were defined according to the default parameters in the standard methodology. FreeBayes, the program used for variant calling in both methodologies, is implemented in Snippy (standard methodology) with a minimum mapping quality of 60, which was used for all the analysis thus far. To evaluate the weight of this parameter on our genome collection when using different references genomes, we decided to decrease the minimum mapping quality value to 20. This value was only tested in the adapted methodology.

The number of SNP differences resulting from this analysis are listed for all DLST types in Table 2 to 4. In general, no major differences were observed by decreasing the minimum mapping quality threshold. For DLST 1-18, the number of SNPs obtained by mapping against the PacBio reference with the minimum mapping quality as 60 was similar to when the value was 20. However, using PA14 as the reference genome with minimum mapping quality 20 increased slightly the number of SNPs between isolates of the same patient. Phylogenetic trees showed also an identical topology, specially for the isolates selected for the comparison (Chapter 7. Supplementary figures). DLST 1-21 isolates were unaltered by a lower minimum mapping quality value as the phylogenetic tree topologies and number of SNPs were very similar between analysis (Chapter 7. Supplementary figures). Interestingly, only changes in genetically distant isolates were detected for DLST 6-7 isolates from 60 to 20 as minimum mapping qualities for mapping performed with both references (Chapter 7. Supplementary figures).

2.3.7.2. Filtering using a lower number of reads per allele to detect a SNP

Another parameter utilized to avoid errors in variant detection is the minimum number of reads per allele to report a SNP. Incorrectly aligned reads can lead to poorly supported genome SNP sites (2). On the adapted methodology, this parameter was part of the filtering applied to the VCF file using an in-house script and the threshold used was 10: a minimum of 10 reads per allele to consider a SNP position. To verify to which extent a stricter value would affect the number of detected SNP differences, we changed the threshold of this parameter to 20.

Only minor decrease in the SNP number was identified when considering sites supported by a minimum of 20 as opposed to 10 reads (Table 2 to 4). Nevertheless, this happened only for genetically distant isolates (Chapter 7. Supplementary figures).

2.4. Discussion

In this study we give insight to the epidemiology of *P. aeruginosa* in the ICUs of the University Hospital of Lausanne by combining a molecular typing method with WGS. While doing so, we also emphasize the importance of defining a bioinformatic pipeline for WGS data analysis suitable for *P. aeruginosa* and the dataset being studied.

MLST results acquired from the isolates' raw reads divided the dataset in three different STs; ST 1076, ST253, and ST17. This division was exactly concordant with the attribution of types performed with DLST; DLST 1-8, DLST 1-21, and DLST 6-7, respectively. Such findings confirmed the previously documented similar discriminatory power of both methods (20). ST253 belongs to the clinical and international well described clonal complex (CC) PA14, and ST17 was previously reported as part of the clonal complex C, both CCs being the worldwide most abundant clonal complexes in the *P. aeruginosa* population (41). No complete published genomes were available at the start

of this study for ST1076 and ST17, although a few investigations have reported these STs (149, 150). Knowing the isolates ST helped to choose the P. aeruginosa PA14 as the reference genome for the mapping step of both methodologies. However, this choice was not ideal as only DLST 1-21 isolates were closely related to the chosen reference.

Applying the standard methodology for WGS analysis, using *P. aeruginosa* PA14 as a reference, lead to an unexpected high number of SNPs between isolates for which epidemiological links were suspected. This happened as well for isolates clustering together in the phylogenetic tree and/or between isolates from the same patient retrieved less than one week apart. These odd results were observed for DLST 1-18 and 6-7. On the other hand, the number of SNPs and the phylogenetic tree topology for DLST cluster 1-21 isolates seemed to be concordant with epidemiological data.

To understand if the SNP problem we encountered was due to an inadequate choice of bioinformatic tools and/or parameters to analyse the WGS data, we adapted the original methodology to a stricter quality approach. A first step of subsampling the reads was added to introduce comparability in coverage which in turn would facilitate the subsequent analysis (147). In addition, more specific recombination and quality filtering was applied. The addition of these analysis steps resulted in a significant decrease of SNP distances between isolates from DLST 1-18 and 6-7. While most of the SNP differences for DLST 1-18 isolates acquired with the standard methodology were between 20 to 100, and 81-121 between isolates of the same patient, the adapted procedure resulted in 0 to 10 between most of the isolates and less than 10 SNPs between isolates of the same patient. The SNP differences between most of the isolates decreased (from 100-200, to 50 to 150), nonetheless it was still in concordance with the sporadic occurrence demonstrated by epidemiological data. The number of SNP differences between isolates of the same patient was now from 7 to 20, instead of 78 to 93. Phylogenetic tree topologies were not majorly changed between approaches as the isolates clustered in an identical manner in both

approaches. No major differences were observed for DLST 1-21 which can be explained by the reference used being from the same ST as this DLST type, and therefore closely related to DLST 1-21 genomes.

SNP analyses is known to give a very high resolution, but a reference genome closely related to the sequenced isolates must be used in order to reduce chances of mismapping and increase regions in the reference genome to which the reads can map against (2). Taking this into consideration, and the results of the previous analyses, we decided to construct a reference genome from the index case of each DLST type and map the reads against it with both methodologies. Interestingly, results of the standard methodology when using the PacBio reference were very similar to the ones obtained when a stricter and more adapted filtering was applied with the *P. aeruginosa* PA14 reference. The decrease in SNP differences was even more evident when the adapted analysis was performed using the PacbBio reference, which produced the lowest number of SNP differences between isolates. Isolates that were distantly related in all methodologies were more affected by the variations in SNP differences. Such findings suggest that the stricter filtering in the adapted methodology is overcoming the problems, i.e. mismapping of reads or lower coverage of certain regions, of mapping against a distantly related reference genome.

Stemming from the premise that filtering options can overcome an inaccurate number of SNP when using a reference not related to the dataset, different parameter thresholds related to the mapping step and site coverage where tested. Changing the mapping qualities to a lower value (MQ 20) resulted in slight changes in SNP differences for DLST 1-18 when analysed with the adapted methodology and *P. aeruginosa* PA14 reference. No significant alterations were detected for the other two DLST types. Similarly, when altering the number of reads from 20 to 10 to consider a site, no major differences in the SNP differences were observed. This implies other parameters or

thresholds are responsible for the differences in results obtained with both methodologies, which need to be further investigated. Standard methodology with other thresholds was not tested so far, but the results could aid in understanding exactly in which parameter does this SNP problem lie. Defining a reliable bioinformatic analysis pipeline adapted to the pathogen and dataset in being studied are important aspects to take in consideration before implementing WGS as a routine epidemiological typing tool (151). Several recent studies on SNP calling bias due to different sequencing and bioinformatic analysis approaches have been published (151, 152). Their findings state the importance for SNP analysis of choosing a close reference strain, as well as departing from good quality raw data, and applying the adequate thresholds, which is comparable to what is reported in this study. As results were equivalent when performing the stricter adapted methodology, this approach was chosen with mapping the reads against the PacBio reference, a MQ of 60, and a minimum number of 10 reads to consider a SNP site, for the subsequent investigation of *P. aeruginosa* epidemiology.

The three investigated DLST types sowed different epidemiological behaviours during this study period. Most of DLST cluster 1-18 patients were hospitalized in the burn unit during overlapping periods of time. As *P. aeruginosa* is capable to survive on wet surfaces such as sinks, sink traps, pipes, and hydrotherapy equipment; several nosocomial outbreaks have been associated with these specific reservoirs (153). DLST 1-18 environmental isolates retrieved from shower mattresses and sink traps from the hydrotherapy room support the assumption of an environmental source of infection. The high number of epidemiological links between patients, along with the wide presence of this DLST type in the environment of the burn unit, helped to previously determine this cluster as responsible for an outbreak with an environmental source (112). Of interest,

the first patient detected was considered as part of the outbreak based on DLST typing, despite no epidemiological links were found to other patients or the environment.

In 2010, DLST type 6-7 was responsible for a small outbreak in the burn unit comprising five patients. From 2011, both DLST 1-21 and 6-7 occurred sporadically throughout the rest of the study period with only one suspected epidemiological link found for DLST 1-21 isolates (between patients and environment). This behaviour may be explained by a major role of this types' prevalence in our ICUs environment, which lead to sporadic patient infection. Nonetheless, one limitation of this study relies on the insufficient environmental sampling information until 2012. A more frequent and regular sampling throughout the four years study would have helped to discover probable epidemiological links between infected patients and environmental sources.

By combining DLST and epidemiological data it was possible to determine three genotypes with different behaviours in our ICUs. However, DLST was not discriminatory enough to confirm possible cases of transmission between patients and between patients and the environment, or to define a probable source of infection. WGS helped to group the DLST 1-18 outbreak isolates with less than 10 SNP differences between them, while excluding Patient 1 as part of the outbreak, which was inferred by epidemiological data but not by DLST typing. Environmental isolates retrieved from sink traps and shower mattresses on the hydrotherapy room clustered with the outbreak isolates (<10 SNPs) which can indicate them as possible sources of infection.

Analysis of DLST 1-21 WGS data confirmed the suspected epidemiological link between isolates retrieved from ICU2. In addition, it considered as closely related, isolates for which no epidemiological links were suspected. For instance, isolates sampled from the burn unit were related with less than 11 SNPs; environmental isolates sampled 10 years apart were related with 11 to 14 SNPs; two isolates from two patients collected 12

years apart had six SNP differences. These values are lower than expected when considering the long time between isolate sampling, and considering that isolates retrieved from the same patient, weeks apart, had close number of SNP differences (0-4 SNPs). One explanation can be the slower evolution of *P. aeruginosa* isolates in the environment of ICUs which then lead to patients being infected with genetically identical strains.

Lastly, a small DLST 6-7 outbreak between patients hospitalized in the burn unit in 2010 was confirmed by WGS (0-13 SNP differences). A subclade of ICU2 clinical and environmental isolates with zero to seven SNP differences suggests a possible transmission between patient and the environment that was not questioned with the epidemiological data. Interestingly, two environmental isolates were associated with long branches. One reason for the occurrence of these long branches is the long branch attraction phenomenon(phylogenetic artefact when distantly related lineages are considered closely related by error because they have both undergone a large amount of molecular change) (154). Another reason could be that these are hypermutator isolates as a response to environmental selection (155). A way to assess the latter would be to investigate the presence of genes coding for the methyl-directed mismatch repair (MMR) system proteins in this DLST type genome.

Although WGS costs are decreasing, its implementation as a routine surveillance method for *P. aeruginosa* still comes at a higher price per isolate than the currently used DLST. Additionally, analyses of WGS data requires a certain level of bioinformatic expertise that is not always available in all epidemiology laboratories(156). Thus, recurring to DLST as a first-line molecular typing tool in combination with the discriminatory power of WGS would culminate in an efficient typing strategy for outbreak investigation or surveillance of *P. aeruginosa* infections.

CHAPTER 3.

High-Quality Complete Genome Sequences of Three *Pseudomonas aeruginosa* Isolates Retrieved from Patients Hospitalized in Intensive Care Units

(*Microbiology Resource Announcements, from the American Society for Microbiology journal. In Press.*)

Bárbara Magalhães,[a] Laurence Senn,[a] and Dominique S. Blanc[a]#

[a]Service of Hospital Preventive Medicine, Lausanne University Hospital, Lausanne, Switzerland

Abstract

Pseudomonas aeruginosa is one of the major Gram-negative pathogens responsible for hospital-acquired infections. Here, we present the high-quality genome sequences of three *P. aeruginosa* genotypes retrieved from patients hospitalized in intensive care units. PacBio reads were assembled into a single contig, which was afterwards corrected using Illumina HiSeq reads.

Pseudomonas aeruginosa is an opportunistic Gram-negative pathogen which is identified as one of the most frequent microorganisms in Intensive Care Units (ICUs) (1, 2).

Following an unexplained increase in P. aeruginosa incidence in the ICUs of the University Hospital of Lausanne, all clinical and environmental isolates from 2010 to 2014 were typed. Most

patients harbored three sequence types (STs): ST1076, ST253, and ST17. To further investigate the epidemiology of this pathogen in the ICUs with short read whole genome sequencing, a complete reference genome was constructed for each ST. The first clinical isolate collected from each of the three STs was selected for that purpose: H25883 (ST1076), H26023 (ST253), and H26027 (ST17).

Single colonies were inoculated in 5ml of Lysogeny broth (LB) and incubated for four hours to reach an early exponential phase. Extraction of the genomic DNA was performed on 1.5 mL cultures using the GenElute bacterial genomic DNA kit (SIGMA-ALDRICH, St. Louis, MO, USA). The genomic DNA (gDNA) was subsequently used for library preparation according to the PacBio standard protocol with the BluePippin size-selection system (Sage Science). The finished libraries were sequenced on a PacBio RSII instrument using P6-C4 chemistry, for 360-min movies, and yielded 100,236 to 103,875 reads with an average size of 19,375 to 19,604 base pairs (bp). Hierarchical Genome Assembly Process (HGAP3) v. 2.3.0 (3) from the SMRT Analysis Software suite (PacBio) was used to assemble the PacBio reads with a minimum seed read length of 6kb. All genomes were manually circularized using the Minimus pipeline (4) included in AMOS (5), merging the overlapping extremities of the main contig. A single circular contig was produced for isolates H25883, H26023, and H26027, with the following genome size and coverage: 6,706,793 (223x), 6,729,215 (217x), and 7,079,586 (228x), respectively.

The extracted gDNA was also used for library preparation with the Nextera DNA Library Preparation Kit (Illumina, Inc., San Diego, CA, USA) for a 100-bp paired-end sequencing on Illumina HiSeq 2500, aiming for a 100-fold coverage. Illumina HiSeq reads were mapped against the assembled PacBio contigs with BWA-MEM, and single nucleotide polymorphisms (SNPs) and indels were identified and corrected using Pilon v.1.22 (6) with a minimum size for unclosed gaps of 10. The genotype, final genome size, and G+C content of the three final corrected circular genomes is represented in Table 5.

A total of 6,400 to 6,806 genes was predicted with Prokaryotic Genome Annotation Pipeline (PGAP) (7), and 6,216 to 6,629 coding sequences (CDSs) annotated, together with 63 to 64 tRNAS and 4 rRNA operons.

Accession number(s). The complete genome sequences for the three *Pseudomonas aeruginosa* isolates have been deposited in DDBJ/ENA/NCBI, and the PacBio and Illumina reads are available in the NCBI Sequence Read Archive. The respective accession numbers are listed in Table 5.

Table 5. Metadata of the three complete corrected genomes of each genotype.

Isolate no.	Genotype	N50 read length (bp)	Genbank accession no.	SRA accession no.		Genome size (bp)	G+C content (%)	CDS
				Illumna reads	PacBio reads			
H25883	ST1076	26,667	CP033686	SRX5329115	SRX5322128	6,706,800	66.15	6,216
H26023	ST253	26,676	CP033685	SRX5329116	SRX5322127	6,729,216	66.21	6,246
H26027	ST17	27,385	CP033684	SRX5329117	SRX5322129	7,079,598	66.07	6,629

Acknowledgments.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

We thank the Ultra-High-Throughput Sequencing (UHTS) unit of the Lausanne Genomic Technologies Facility (LGTF) at the University of Lausanne for PacBio and Illumina HiSeq sequencing services. We also thank the members of the LGTF Bioinformatics unit for the genomes assembly and assistance with post-assembly analysis.

References (Chapter 6. References)

CHAPTER 4.

Comparison of different bioinformatic approaches for routine analyses of WGS data

4.1. Objectives

Whole genome sequencing (WGS) has proven to be a very effective subtyping tool for various nosocomial pathogens. An array of approaches to analyse WGS data have been applied for epidemiologic and infection control purposes. The genomic data can be exploited essentially by single nucleotide polymorphisms (SNPs) or gene-by-gene methods. Only a few studies were published to date on the comparison of these different analysis approaches, leaving out several important pathogens. A previously published ST228 Methicillin-resistant *Staphylococcus aureus* (MRSA) dataset from a large outbreak in a University hospital (1) will be used to assess and compare the performance of different genomic methods for outbreak investigation. This data set was investigated through mapping against a reference and posterior SNP calling and showed the outbreak clonality with a diversification through time of seven different branches. With this project, the same dataset will be analysed with Whole Genome SNPs (wgSNPs) and Whole Genome Multi-Locus Sequence Typing (wgMLST), both implemented in BioNumerics v.7.6.3

(Applied-Maths). Ultimately, the comparison of these methodologies will help to evaluate their implementation in routine diagnostic of *S. aureus*.

4.2. Material and Methods

4.2.1. Bacterial isolates

All MRSA ST 228 isolates (one per patient) recovered in 2008, and one consecutive patient out of 10 from 2009 to 2012, were included in the study.

4.2.2. Whole genome SNPs analysis

Reads of 235 MRSA sequence type 228 (ST 228) isolates were imported into BioNumerics. *Staphylococcus aureus subsp. aureus* N315 reference strain was used as the reference sequence for mapping against the raw reads. Base differences against the reference sequence were obtained by the wgSNPs tool using the "Strict SNP filtering (closed SNP set)" option. This option removes positions with at least one unreliable base (N), ambiguous base (non-ATCG), gaps and non-discriminatory positions between the entries. Each SNP position should had at least 5x coverage, was covered once in forward or reverse direction, and the minimum inter-SNP distance was 12. A maximum likelihood tree cannot be obtained with BioNumerics as it considers nucleotides as characters. Thus, the filtered SNPs matrix was exported to the comparison window where a maximum spanning tree (MST) was constructed.

4.2.3. Whole genome MLST analysis

Whole genome MLST analysis was done by using the default parameters suggested for S*. aureus*. The wgMLST pan-genome scheme for *S. aureus* includes a total of 3897 loci, from which 1861 are considered in the core genome, and 2036 as accessory loci. Three jobs were submitted to the calculation engine of Amazon: *de novo* assembly, assembly-

free calls, and assembly-based calls. Alleles were identified by combining the assembly-free k-mer based approach from raw reads and the assembly-based BLAST approach of Velvet optimizer assembled genomes. New alleles were automatically submitted to the allele nomenclature server. A maximum spanning tree (MST) was constructed for the wgMLST and cgMLST allelic profile.

4.3. Results

4.3.1. Whole genome SNPs results

Figure 36 shows the maximum likelihood tree previously published by SNPs analysis with all MRSA ST 228 isolates. The isolates' phylogeny based on the SNP variable sites illustrated the diversification of seven major branches during the 52-month study period. Isolates considered not part of the outbreak were distantly related to the outbreak isolates. In addition, isolate number 188 (group b) was distantly related to the other members of the group and was represented by a long branch.

Whole genome SNPs analysis in BioNumerics resulted in a total of 879 SNPs between the 235 MRSA isolates. In general, isolates belonging to a specific branch group (a to g) tended to cluster together with few SNP differences, with the exception of some isolates belonging to group b, d, and g, in red, pink and dark blue, respectively (Figure 37). Isolate 188 was closely related to the remaining members of its group (group b). Isolates not related to the outbreak were distantly related from the remaining isolates, with high number of SNPs between them (>38 SNPs). Isolate 233, previously considered the most distant isolate, showed the highest number of SNP differences (87 SNPs) with the remaining isolates.

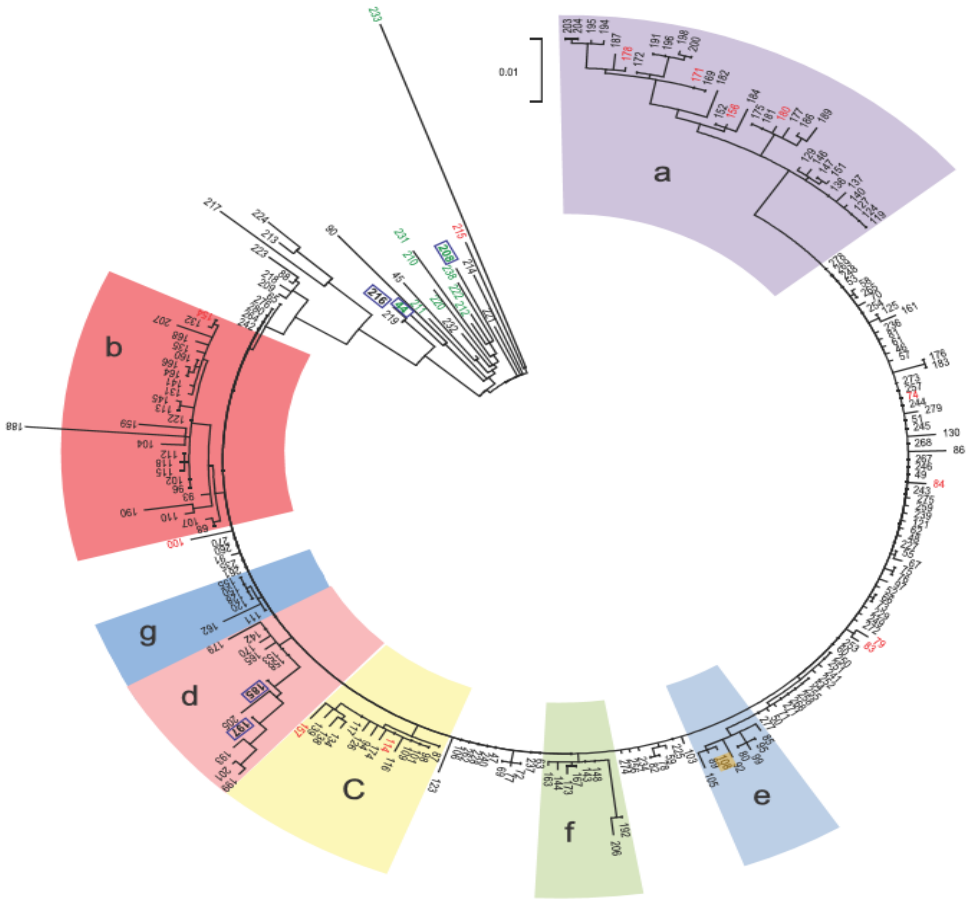


Figure 36. Maximum likelihood tree based on SNP variable sites of all S. aureus ST 228 isolates over a 52-month period. Seven different branches are highlighted in different colors and numerated from a to g (1).

4.3.2. Whole genome MLST results

Results were not in congruence with the SNPs analysis approach (Figure 38). First, isolates obtained before the outbreak clustered within the outbreak isolates. Second, the clustering of the isolates in the seven branches (a to g groups) was not observed; they were mixed in different clusters. Third, isolate 233, which was considered as the more distant with the SNPs analysis, was found to cluster with the remaining isolates. In addition, isolate 188 was closely related to other isolates and not located at the end of a long branch. In turn, isolate 108 showed high allele differences from the remaining isolates (2623 alleles), as well as isolates 223 and 221 (2565 alleles).

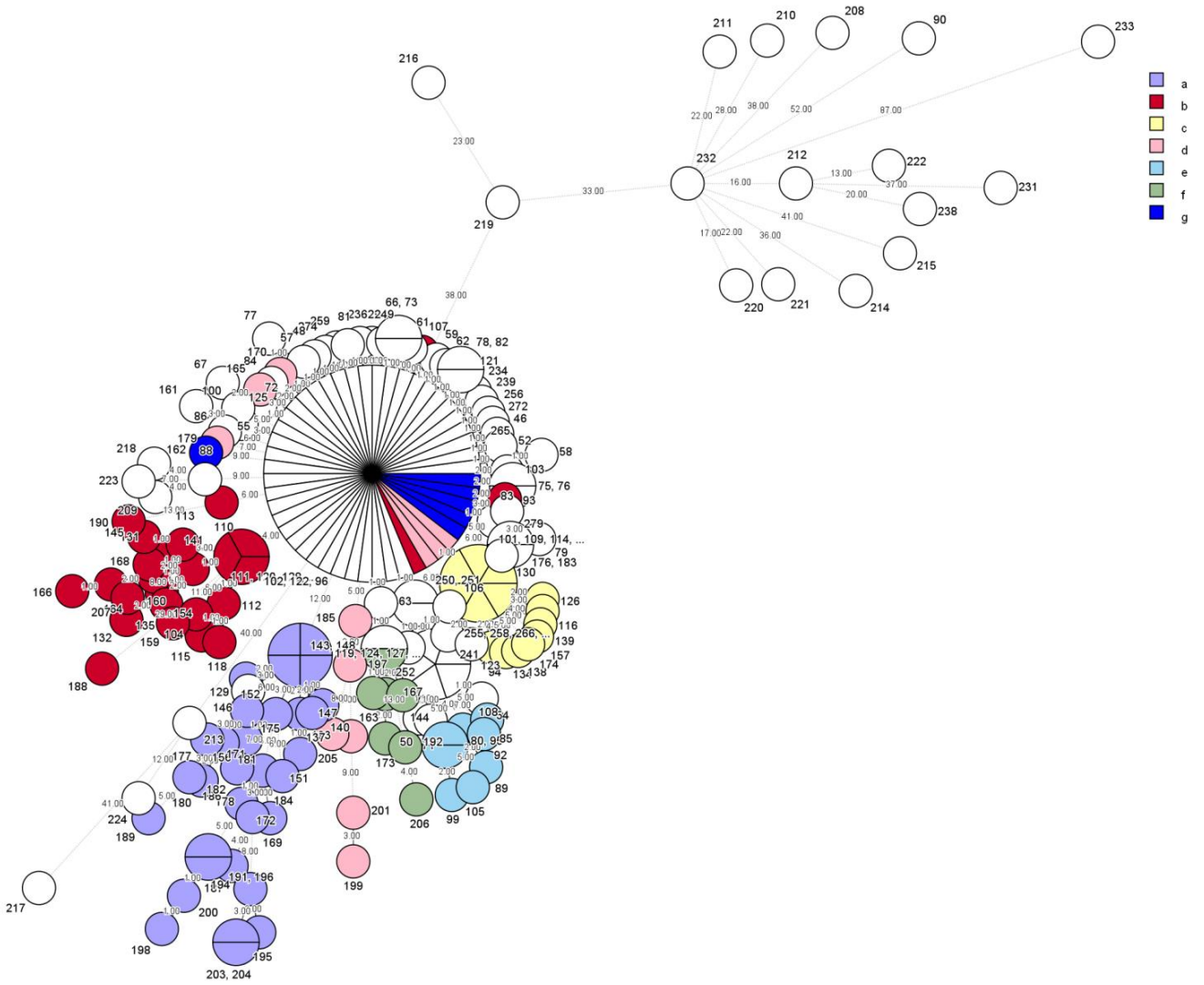


Figure 37. Minimum spanning tree based on *S. aureus* ST 228 isolates' SNP differences acquired with wgSNPs. SNPs differences are discriminated on the branches of tree. Isolates belonging to the seven previously reported clusters are highlighted in different colors

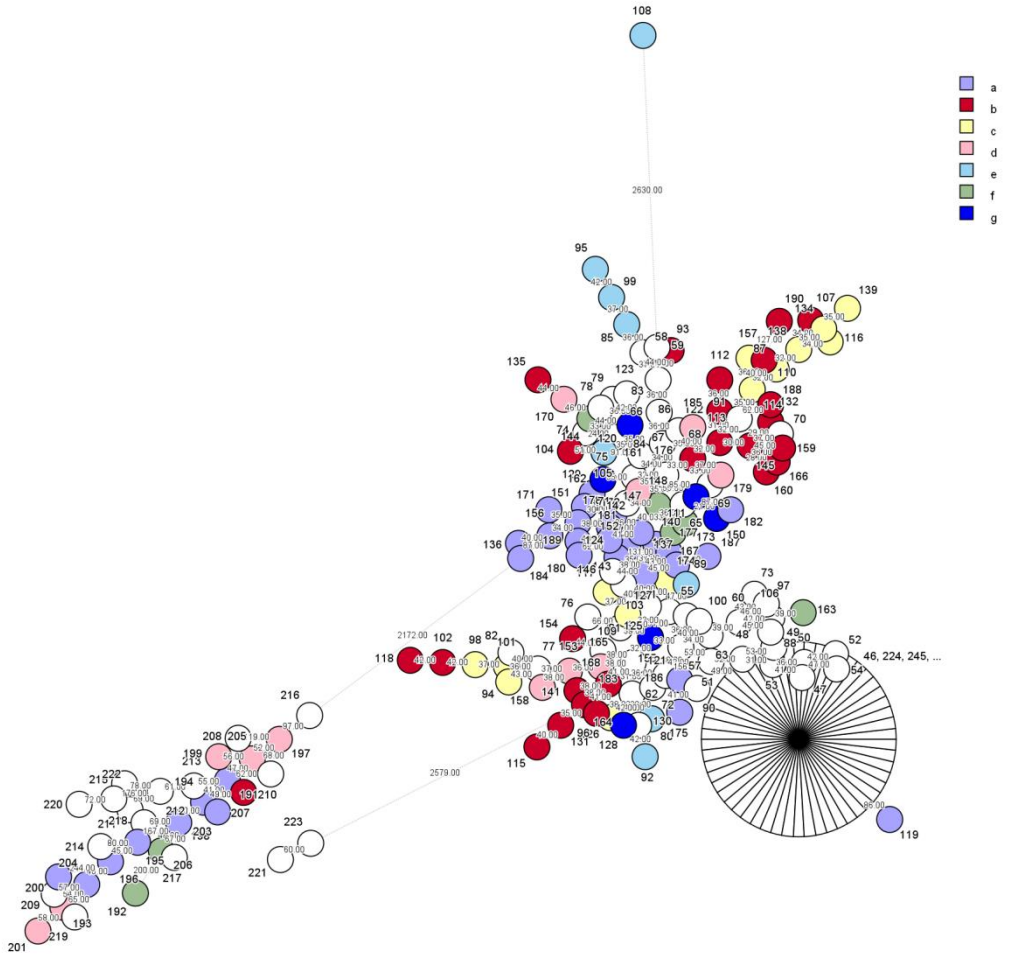


Figure 38. Minimum spanning tree based on *S. aureus* ST 228 isolates' allele differences acquired with wgMLST. Isolates belonging to the seven previously reported clusters are highlighted in different colors

4.3.3. Whole genome MLST results after bug fix

While trying to decipher why discrepancies in the results were obtained with wgSNPs and wgMLST, we discovered that low quality alleles were being marked as tentative, i.e. does not meet the quality criteria set by the curator of the database and are not part of the search data for the allele calling algorithms. However, these alleles were being included in the search data by the allele calling algorithms. After discussion with the BioNumerics support team on this matter, we realized the problem relied on the corruption of the database with low quality alleles as users were submitting alleles with very low similarity thresholds and the program did not check the start, stop, and internal stop was performed. Consequently, some loci fragments were being considered as allele. To overcome this problem, BioNumerics team removed all tentative alleles from the search data. Although they are still in the allele nomenclature database, they are no longer used as input for allele calling algorithms.

wgMLST was repeated only for a subset of MRSA ST228 isolates. From the original 235 isolates, 131 were selected to be analysed by the corrected version of wgMLST. Figure 39 shows the concordance between new wgMLST results and the ones obtained with wgSNPs and the original SNPs analysis. Clustering of isolates according to the previously described groups was more accurate than the one obtained with wgSNPs. Isolates from group f (in brown) evolved from the ancestor group of isolates in green but some were disposed separately on the MST tree. Isolates 233 and 221 were part of the non-outbreak group in white whit 82 and 15 allele differences,

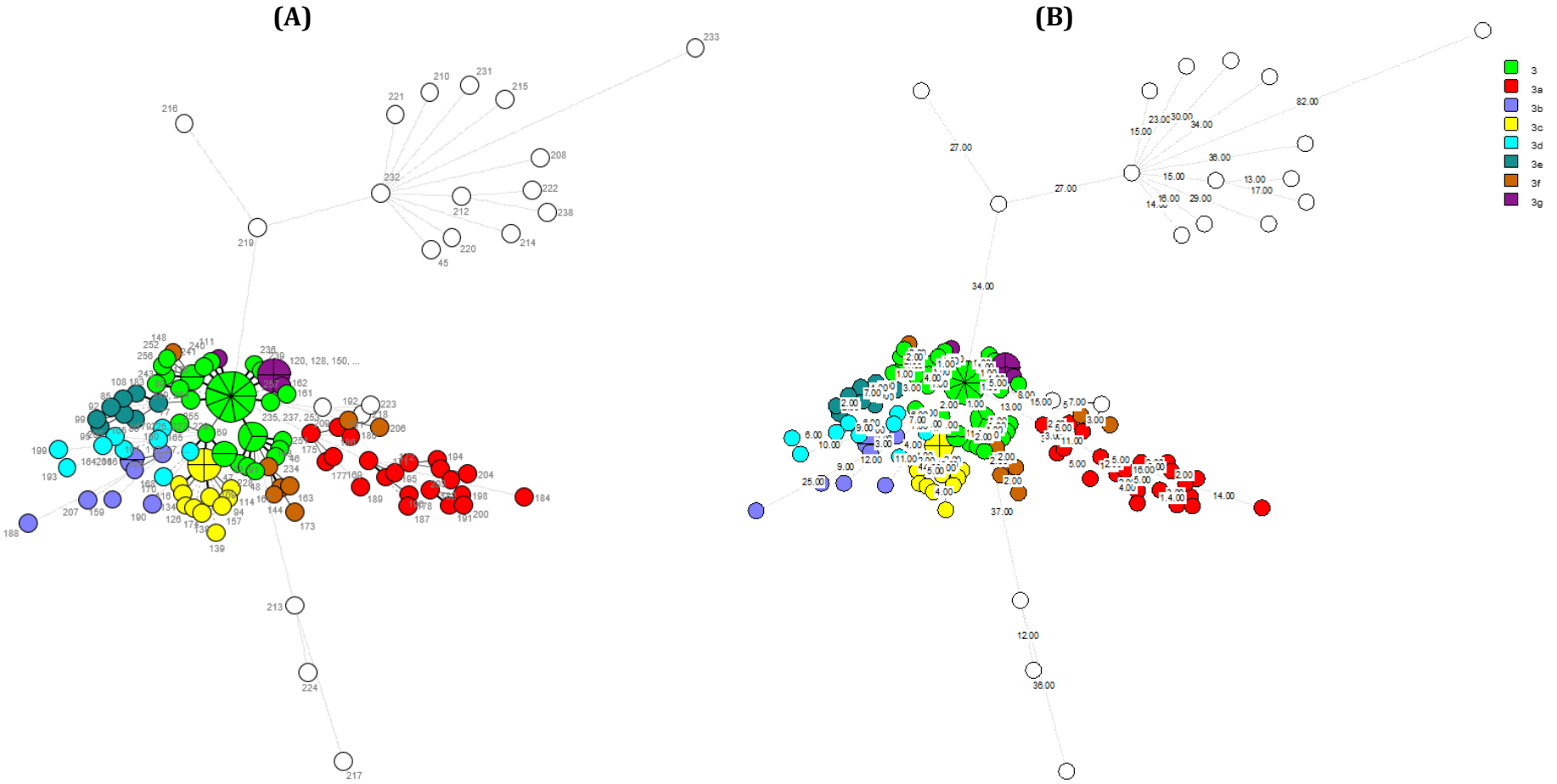


Figure 39. Minimum spanning tree based ona subset of 131 *S. aureus* ST 228 isolates' allele differences acquired with wgMLST after the bug fix. Two MST trees are illustrated: (A) with isolate number and (B) with allele differences between isolates. Isolates belonging to the seven previously reported clusters are highlighted in different colors

respectively. Isolate 188 clustered with isolates from group b (in blue) with 25 allele differences and isolate 108 belonged to group e (in dark green) with seven allele differences.

4.4. Discussion

A great concern with WGS in a clinical routine microbiology laboratory is still data analysis, as it normally requires specific bioinformatic expertise. SNP and gene-by-gene comparisons are the primary approaches for WGS data analyses for purposes of bacterial strain typing and epidemiological investigations (2). However, due to the availability of these two approaches in user-friendly software packages, WGS data analysis for diagnostic purposes is possible with little bioinformatics knowledge (25). This project aimed to compare SNP analysis and gene-by-gene approaches for WGS data analysis in routine investigation of *S. aureus*.

Whole genome SNPs analysis performed with BioNumerics was able to cluster the isolates similarly to what was observed in the original published results, with only some exceptions. The number of SNP positions was lower than the one previously published (879 vs1565 SNP positions), which can maybe be explained by the application of a stricter filtering of SNPs. After the bug fix, wgMLST results were concordant with both SNP analysis methods. The subset of 131 isolates clustered accordingly to the seven branches almost identically to the original phylogenetic tree. Similar distances between isolates were found between wgSNPs and wgMLST, e.g. Patient 233 was the farthest from the outbreak isolates with 87 SNPs and 82 allele differences, respectively. Unfortunately, the complete dataset was not included in the repetition of wgMLST since BioNumerics requires the payment of credits to perform the analysis with wgMLST tools. Although the concordance in SNP and allele differences between methods seems evident, other *S.*

aureus datasets should be tested in order to validate BioNumerics tools for outbreak investigation of this pathogen.

Important aspects of these two approaches need to be considered when deciding which method to apply to the dataset in question. SNP analysis has high resolution but a closely related genome reference to the isolates is crucial for correct calling of variant sites and that is not always possible for real-time outbreak investigation. Another limitation of this approach is the reproducibility between different studies where different references and thresholds are being applied. Whole genome MLST is able to surpass these obstacles as the mapping step is not performed, hence there is no need to choose a closely related reference. In addition, the assignment of alleles in comparison to a curated set of predefined genes gives it interlaboratory reproducibility (2).

In general, both BioNumerics WGS tools were very easy to use, in relatively short amount of time. No knowledge on bioinformatics was needed to perform these BioNumerics analyses, as it could be easily followed by the software tutorials. However, such knowledge is greatly needed in the investigation of discrepancies between these methods. For instance, *P. aeruginosa* is not a clonal pathogen as opposed to *S. aureus*, and the default parameters applied to one pathogen may not work best for another. Understanding the microorganism in question, which parameters and thresholds should be altered, is still crucial for a reliable application of such software to diagnostics.

CHAPTER 5.

Conclusions and Future Perspectives

The main objective of this thesis was to evaluate the implementation of WGS for epidemiological investigation of nosocomial pathogens. To accomplish that, several WGS analysis methodologies tested using datasets from two important nosocomial pathogens, *P. aeruginosa* and MRSA.

DLST grouping of the isolates in three types, DLST 1-18, DLST 1-21, and DLST 6-7, was identical to ST identification: ST1076, ST253, and ST17, respectively. Applying a standard methodology with default parameters, while using a reference distantly related to the isolate collection, lead to an unexpected high number of SNP differences between epidemiologically linked isolates. A stricter quality filtering helped to overcome this problem. When using a closely related reference, the results were very similar to the ones obtained with stricter quality thresholds. This indicates that mapping against a distant reference creates variant artefacts due to the low quality of the alignment. However, when parameters related to alignment quality were changed, no significant changes in the number of SNPs were observed. Using more discrepant thresholds or evaluating other important parameters could aid in understanding where the odd SNPs number problem resides. Additionally, the changed thresholds need to be tested with the standard methodology as well.

After choosing the methodology that seemed more reliable in results and in principle for *P. aeruginosa* investigation, the obtained WGS data were highly concordant with the epidemiological information. WGS confirmed the occurrence of a long DLST 1-18 outbreak, and a DLST 6-7 smaller one, in the burn unit. It successfully identified genetically related isolates for which epidemiological links were observed. In turn, WGS considered closely related isolates that were not epidemiologically linked, suggesting that the environment could play an important role in the sporadic infection of patients, as well as the source of nosocomial outbreaks.

Due to the costs and need of bioinformatic expertise when using WGS routinely in epidemiology laboratories, combining DLST as a first screening method and complementing it with the discriminatory power of WGS to resolve specific cases looks like a very efficient and reliable typing approach.

One of this project's tasks was to genetically characterize each DLST type; however this was not accomplished so far. Preliminary results on virulence and resistance genes have shown that DLST 1-18 and DLST 6-7 have comparable virulence and resistance. Although DLST 1-21 resistance does not differ from the other types, its virulence arsenal appears to be bigger. Comparison of the PacBio references of each DLST type will further elucidate on the genetic composition of each genotype.

It is known that a single reference genome is not sufficient to fully represent the entire genetic diversity of a given species (161). Thus, the analysis of these *P. aeruginosa* isolates' pangenome will be performed in order to determined its core and variable/accessory/dispensable gene content.

The second project aimed to compare different approaches for WGS data analysis, using both open-access and commercially available bioinformatic tools. Only the wgSNPs

and wgMLST tools incorporated in BioNumerics program were tested thus far. These tools showed high concordance in tree topology and genetic distances with the SNPs analysis previously performed. In addition, wgMLST showed great potential to be used as an alternative approach to SNP analysis for routine molecular typing of nosocomial pathogens. However, adaptability of the analysis to the pathogen and dataset being studied is not easily performed with these commercial programs since the parameters are limited.

In order to validate wgMLST for its implementation in epidemiology laboratories, other MRSA datasets and pathogens needed to be tested and the results evaluated. Investigation of other wgMLST-based tools, such as SeqSphere+ (162), whole genome sequence analysis (WGSA) (https://pathogen.watch/), and PHYLOViZ (163), would greatly complement this comparison study and put in evidence the reproducibility of WGS analysis.

CHAPTER 6.

References

1. Senn L, Clerc O, Zanetti G, Basset P, Prod'hom G, Gordon NC, Sheppard AE, Crook DW, James R, Thorpe HA, Feil EJ, Blanc DS. 2016. The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant *Staphylococcus aureus*. MBio 7:e02039-15.
2. Schurch AC, van Schaik W. 2017. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. Ann N Y Acad Sci 1388:108-120.
3. Sabat AJ, Budimir A, Nashev D, Sa-Leao R, van Dijl J, Laurent F, Grundmann H, Friedrich AW, Markers ESGoE. 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill 18:20380.
4. MacCannell D. 2013. Bacterial strain typing. Clin Lab Med 33:629-50.
5. Losada-Perez M, Gabilondo H, Molina I, Turiegano E, Torroja L, Thor S, Benito-Sipos J. 2013. Klumpfuss controls FMRFamide expression by enabling BMP signaling within the NB5-6 lineage. Development 140:2181-9.
6. Francino MP. 2012. The ecology of bacterial genes and the survival of the new. Int J Evol Biol 2012:394026.
7. Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol 2:519-23.
8. Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12:263-73.
9. Gurtler V, Mayall BC. 2001. Genomic approaches to typing, taxonomy and evolution of bacterial isolates. Int J Syst Evol Microbiol 51:3-16.
10. van Belkum A. 2003. High-throughput epidemiologic typing in clinical microbiology. Clin Microbiol Infect 9:86-100.
11. Du XF, Xiao M, Liang HY, Sun Z, Jiang YH, Chen GY, Meng XY, Zou GL, Zhang L, Liu YL, Zhang H, Sun HL, Jiang XF, Xu YC. 2014. An improved MLVF method and its comparison with traditional MLVF, spa typing, MLST/SCCmec and PFGE for the typing of methicillin-resistant *Staphylococcus aureus*. Int J Mol Sci 15:725-42.
12. Sharma-Kuinkel BK, Rude TH, Fowler VG, Jr. 2016. Pulse Field Gel Electrophoresis. Methods Mol Biol 1373:117-30.

13. Scholz CF, Jensen A. 2017. Development of a Single Locus Sequence Typing (SLST) Scheme for Typing Bacterial Species Directly from Complex Communities. Methods Mol Biol 1535:97-107.
14. Maiden MC. 2006. Multilocus sequence typing of bacteria. Annu Rev Microbiol 60:561-88.
15. Ranjbar R, Karami A, Farshad S, Giammanco GM, Mammina C. 2014. Typing methods used in the molecular epidemiology of microbial pathogens: a how-to guide. New Microbiol 37:1-15.
16. Basset P, Hammer NB, Kuhn G, Vogel V, Sakwinska O, Blanc DS. 2009. *Staphylococcus aureus clfB* and *spa* alleles of the repeat regions are segregated into major phylogenetic lineages. Infect Genet Evol 9:941-7.
17. Basset P, Senn L, Prod'hom G, Bille J, Francioli P, Zanetti G, Blanc DS. 2010. Usefulness of double locus sequence typing (DLST) for regional and international epidemiological surveillance of methicilin-resistant *Staphylococcus aureus*. Clin Microbiol Infect 16:1289-96.
18. Basset P, Blanc DS. 2014. Fast and simple epidemiological typing of *Pseudomonas aeruginosa* using the double-locus sequence typing (DLST) method. Eur J Clin Microbiol Infect Dis 33:927-32.
19. Sakwinska O, Blanc DS, Lazor-Blanchet C, Moreillon M, Giddey M, Moreillon P. 2010. Ecological temporal stability of *Staphylococcus aureus* nasal carriage. J Clin Microbiol 48:2724-8.
20. Cholley P, Stojanov M, Hocquet D, Thouverez M, Bertrand X, Blanc DS. 2015. Comparison of double-locus sequence typing (DLST) and multilocus sequence typing (MLST) for the investigation of *Pseudomonas aeruginosa* populations. Diagn Microbiol Infect Dis 82:274-7.
21. Human Genome Sequencing C. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931-945.
22. Grada A, Weinbrecht K. 2013. Next-generation sequencing: methodology and application. J Invest Dermatol 133:e11.
23. Moran-Gilad J. 2017. Whole genome sequencing (WGS) for food-borne pathogen surveillance and control - taking the pulse. Euro Surveill 22.
24. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S, Kooistra-Smid AM, Raangs EC, Rosema S, Veloo AC, Zhou K, Friedrich AW, Rossen JW. 2017. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 243:16-24.
25. Kwong JC, McCallum N, Sintchenko V, Howden BP. 2015. Whole genome sequencing in clinical and public health microbiology. Pathology 47:199-210.
26. Allard MW. 2016. The Future of Whole-Genome Sequencing for Public Health and the Clinic. J Clin Microbiol 54:1946-8.
27. Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A, Galaxy T. 2011. Integrating diverse databases into an unified analysis framework: a Galaxy approach. Database (Oxford) 2011:bar011.
28. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM. 2014. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 58:212-20.
29. Bergey DH, Holt JG. 1994. Bergey's manual of determinative bacteriology. Williams & Wilkins, Baltimore.
30. Spiers AJ, Buckling A, Rainey PB. 2000. The causes of *Pseudomonas* diversity. Microbiology 146 (Pt 10):2345-50.

31. Silby MW, Winstanley C, Godfrey SA, Levy SB, Jackson RW. 2011. *Pseudomonas* genomes: diverse and adaptable. FEMS Microbiol Rev 35:652-80.
32. Blanc DS, Francioli P, Zanetti G. 2007. Molecular Epidemiology of *Pseudomonas aeruginosa* in the Intensive Care Units - A Review. Open Microbiol J 1:8-11.
33. Valentini M, Storelli N, Lapouge K. 2011. Identification of C(4)-dicarboxylate transport systems in *Pseudomonas aeruginosa* PAO1. J Bacteriol 193:4307-16.
34. Ornston LN. 1971. Regulation of catabolic pathways in *Pseudomonas.* Bacteriol Rev 35:87-116.
35. Sias SR, Stouthamer AH, Ingraham JL. 1980. The assimilatory and dissimilatory nitrate reductases of *Pseudomonas aeruginosa* are encoded by different genes. J Gen Microbiol 118:229-34.
36. Murray JL, Kwon T, Marcotte EM, Whiteley M. 2015. Intrinsic Antimicrobial Resistance Determinants in the Superbug *Pseudomonas aeruginosa*. MBio 6:e01603-15.
37. Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R, McLaughlin SF, Peckham HE, Omberg L, Bormann Chung CA, Stanley S, Pearlstein K, Levandowsky E, Acevedo-Acevedo S, Auton A, Keinan A, Acuna-Alonzo V, Barquera-Lozano R, Canizales-Quinteros S, Eng C, Burchard EG, Russell A, Reynolds A, Clark AG, Reese MG, Lincoln SE, Butte AJ, De La Vega FM, Bustamante CD. 2012. Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am J Hum Genet 91:660-71.
38. Maatallah M, Cheriaa J, Backhrouf A, Iversen A, Grundmann H, Do T, Lanotte P, Mastouri M, Elghmati MS, Rojo F, Mejdi S, Giske CG. 2011. Population structure of *Pseudomonas aeruginosa* from five Mediterranean countries: evidence for frequent recombination and epidemic occurrence of CC235. PLoS One 6:e25617.
39. Pirnay JP, Bilocq F, Pot B, Cornelis P, Zizi M, Van Eldere J, Deschaght P, Vaneechoutte M, Jennes S, Pitt T, De Vos D. 2009. *Pseudomonas aeruginosa* population structure revisited. PLoS One 4:e7740.
40. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE, Lory S, Olson MV. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406:959-64.
41. Cramer N, Wiehlmann L, Ciofu O, Tamm S, Hoiby N, Tummler B. 2012. Molecular epidemiology of chronic *Pseudomonas aeruginosa* airway infections in cystic fibrosis. PLoS One 7:e50731.
42. Mathee K, Narasimhan G, Valdes C, Qiu X, Matewish JM, Koehrsen M, Rokas A, Yandava CN, Engels R, Zeng E, Olavarietta R, Doud M, Smith RS, Montgomery P, White JR, Godfrey PA, Kodira C, Birren B, Galagan JE, Lory S. 2008. Dynamics of *Pseudomonas aeruginosa* genome evolution. Proc Natl Acad Sci U S A 105:3100-5.
43. Xavier DE, Picao RC, Girardello R, Fehlberg LC, Gales AC. 2010. Efflux pumps expression and its association with porin down-regulation and beta-lactamase production among *Pseudomonas aeruginosa* causing bloodstream infections in Brazil. BMC Microbiol 10:217.
44. Kung VL, Ozer EA, Hauser AR. 2010. The accessory genome of *Pseudomonas aeruginosa*. Microbiol Mol Biol Rev 74:621-41.
45. Ozer EA, Allen JP, Hauser AR. 2014. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. BMC Genomics 15:737.

46. Driscoll JA, Brody SL, Kollef MH. 2007. The epidemiology, pathogenesis and treatment of *Pseudomonas aeruginosa* infections. Drugs 67:351-68.
47. Sadikot RT, Blackwell TS, Christman JW, Prince AS. 2005. Pathogen-host interactions in *Pseudomonas aeruginosa* pneumonia. Am J Respir Crit Care Med 171:1209-23.
48. Trouillet JL, Vuagnat A, Combes A, Kassis N, Chastre J, Gibert C. 2002. *Pseudomonas aeruginosa* ventilator-associated pneumonia: comparison of episodes due to piperacillin-resistant versus piperacillin-susceptible organisms. Clin Infect Dis 34:1047-54.
49. Nicotra MB, Rivera M, Dale AM, Shepherd R, Carter R. 1995. Clinical, pathophysiologic, and microbiologic characterization of bronchiectasis in an aging cohort. Chest 108:955-61.
50. Chatzinikolaou I, Abi-Said D, Bodey GP, Rolston KV, Tarrand JJ, Samonis G. 2000. Recent experience with *Pseudomonas aeruginosa* bacteremia in patients with cancer: Retrospective analysis of 245 episodes. Arch Intern Med 160:501-9.
51. Hamasuna R, Betsunoh H, Sueyoshi T, Yakushiji K, Tsukino H, Nagano M, Takehara T, Osada Y. 2004. Bacteria of preoperative urinary tract infections contaminate the surgical fields and develop surgical site infections in urological operations. Int J Urol 11:941-7.
52. Mesaros N, Nordmann P, Plesiat P, Roussel-Delvallez M, Van Eldere J, Glupczynski Y, Van Laethem Y, Jacobs F, Lebecque P, Malfroot A, Tulkens PM, Van Bambeke F. 2007. *Pseudomonas aeruginosa*: resistance and therapeutic options at the turn of the new millennium. Clin Microbiol Infect 13:560-78.
53. Lyczak JB, Cannon CL, Pier GB. 2000. Establishment of *Pseudomonas aeruginosa* infection: lessons from a versatile opportunist. Microbes Infect 2:1051-60.
54. Gellatly SL, Hancock RE. 2013. *Pseudomonas aeruginosa*: new insights into pathogenesis and host defenses. Pathog Dis 67:159-73.
55. Kurahashi K, Kajikawa O, Sawa T, Ohara M, Gropper MA, Frank DW, Martin TR, Wiener-Kronish JP. 1999. Pathogenesis of septic shock in *Pseudomonas aeruginosa* pneumonia. J Clin Invest 104:743-50.
56. O'Toole GA, Kolter R. 1998. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. Mol Microbiol 30:295-304.
57. He SJ, Jin YM, Huang AR, Wang CX, Zhou AH, Wang X, Shan XO. 2008. [Clinical analysis of community-acquired *Pseudomonas aeruginosa* septic shock]. Zhonghua Er Ke Za Zhi 46:333-9.
58. Beatson SA, Whitchurch CB, Semmler AB, Mattick JS. 2002. Quorum sensing is not required for twitching motility in *Pseudomonas aeruginosa.* J Bacteriol 184:3598-604.
59. Sriramulu DD, Lunsdorf H, Lam JS, Romling U. 2005. Microcolony formation: a novel biofilm model of *Pseudomonas aeruginosa* for the cystic fibrosis lung. J Med Microbiol 54:667-76.
60. King JD, Kocincova D, Westman EL, Lam JS. 2009. Review: Lipopolysaccharide biosynthesis in *Pseudomonas aeruginosa*. Innate Immun 15:261-312.
61. Kipnis E, Sawa T, Wiener-Kronish J. 2006. Targeting mechanisms of *Pseudomonas aeruginosa* pathogenesis. Med Mal Infect 36:78-91.
62. Deep A, Chaudhary U, Gupta V. 2011. Quorum sensing and Bacterial Pathogenicity: From Molecules to Disease. J Lab Physicians 3:4-11.

63. Schuster M, Lostroh CP, Ogi T, Greenberg EP. 2003. Identification, timing, and signal specificity of *Pseudomonas aeruginosa* quorum-controlled genes: a transcriptome analysis. J Bacteriol 185:2066-79.
64. Bjarnsholt T, Tolker-Nielsen T, Hoiby N, Givskov M. 2010. Interference of *Pseudomonas aeruginosa* signalling and biofilm formation for infection control. Expert Rev Mol Med 12:e11.
65. Lieleg O, Caldara M, Baumgartel R, Ribbeck K. 2011. Mechanical robustness of *Pseudomonas aeruginosa* biofilms. Soft Matter 7:3307-3314.
66. Wei Q, Ma LZ. 2013. Biofilm matrix and its regulation in *Pseudomonas aeruginosa.* Int J Mol Sci 14:20983-1005.
67. Sawa T, Wiener-Kronish JP. 2004. A therapeutic strategy against the shared virulence mechanism utilized by both Yersinia pestis and *Pseudomonas aeruginosa*. Anesthesiol Clin North America 22:591-606, viii-ix.
68. Engel J, Balachandran P. 2009. Role of *Pseudomonas aeruginosa* type III effectors in disease. Curr Opin Microbiol 12:61-6.
69. Hauser AR. 2009. The type III secretion system of *Pseudomonas aeruginosa:* infection by injection. Nat Rev Microbiol 7:654-65.
70. Davies JC. 2002. *Pseudomonas aeruginosa* in cystic fibrosis: pathogenesis and persistence. Paediatr Respir Rev 3:128-34.
71. Giamarellou H. 2002. Prescribing guidelines for severe *Pseudomonas* infections. J Antimicrob Chemother 49:229-33.
72. Tumbarello M, Repetto E, Trecarichi EM, Bernardini C, De Pascale G, Parisini A, Rossi M, Molinari MP, Spanu T, Viscoli C, Cauda R, Bassetti M. 2011. Multidrug-resistant *Pseudomonas aeruginosa* bloodstream infections: risk factors and mortality. Epidemiol Infect 139:1740-9.
73. Hirsch EB, Tam VH. 2010. Impact of multidrug-resistant *Pseudomonas aeruginosa* infection on patient outcomes. Expert Rev Pharmacoecon Outcomes Res 10:441-51.
74. Breidenstein EB, de la Fuente-Nunez C, Hancock RE. 2011. *Pseudomonas aeruginosa*: all roads lead to resistance. Trends Microbiol 19:419-26.
75. Morita Y, Tomida J, Kawamura Y. 2014. Responses of *Pseudomonas aeruginosa* to antimicrobials. Front Microbiol 4:422.
76. Skiada A, Markogiannakis A, Plachouras D, Daikos GL. 2011. Adaptive resistance to cationic compounds in *Pseudomonas aeruginosa*. Int J Antimicrob Agents 37:187-93.
77. Fernandez L, Breidenstein EB, Song D, Hancock RE. 2012. Role of intracellular proteases in the antibiotic resistance, motility, and biofilm formation of *Pseudomonas aeruginosa*. Antimicrob Agents Chemother 56:1128-32.
78. Poole K. 2011. *Pseudomonas aeruginosa*: resistance to the max. Front Microbiol 2:65.
79. Pfeifer Y, Cullik A, Witte W. 2010. Resistance to cephalosporins and carbapenems in Gram-negative bacterial pathogens. Int J Med Microbiol 300:371-9.
80. Poole K. 2005. Efflux-mediated antimicrobial resistance. J Antimicrob Chemother 56:20-51.
81. Zilberberg MD, Chen J, Mody SH, Ramsey AM, Shorr AF. 2010. Imipenem resistance of *Pseudomonas* in pneumonia: a systematic literature review. BMC Pulm Med 10:45.

82. Drlica K, Hiasa H, Kerns R, Malik M, Mustaev A, Zhao X. 2009. Quinolones: action and resistance updated. Curr Top Med Chem 9:981-98.
83. Jacoby GA. 2005. Mechanisms of resistance to quinolones. Clin Infect Dis 41 Suppl 2:S120-6.
84. Taccetti G, Campana S, Neri AS, Boni V, Festini F. 2008. Antibiotic therapy against *Pseudomonas aeruginosa* in cystic fibrosis. J Chemother 20:166-9.
85. Poole K. 2005. Aminoglycoside resistance in *Pseudomonas aeruginosa*. Antimicrob Agents Chemother 49:479-87.
86. Control ECfDPa. 2017. Antimicrobial resistance surveillance in Europe 2015. Stockholm: ECDC.
87. Blanc DS. 2004. The use of molecular typing for epidemiological surveillance and investigation of endemic nosocomial infections. Infect Genet Evol 4:193-7.
88. Gautom RK. 1997. Rapid pulsed-field gel electrophoresis protocol for typing of *Escherichia coli* O157:H7 and other gram-negative organisms in 1 day. J Clin Microbiol 35:2977-80.
89. Salimi J. 2009. On the management of mycotic femoral pseudoaneurysms in intravenous drug abusers. Ann Vasc Surg 23:824.
90. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233-9.
91. Botes J, Williamson G, Sinickas V, Gurtler V. 2003. Genomic typing of *Pseudomonas aeruginosa* isolates by comparison of Riboprinting and PFGE: correlation of experimental results with those predicted from the complete genome sequence of isolate PAO1. J Microbiol Methods 55:231-40.
92. Sobral D, Mariani-Kurkdjian P, Bingen E, Vu-Thien H, Hormigos K, Lebeau B, Loisy-Hamon F, Munck A, Vergnaud G, Pourcel C. 2012. A new highly discriminatory multiplex capillary-based MLVA assay as a tool for the epidemiological survey of *Pseudomonas aeruginosa* in cystic fibrosis patients. Eur J Clin Microbiol Infect Dis 31:2247-56.
93. Vu-Thien H, Corbineau G, Hormigos K, Fauroux B, Corvol H, Clement A, Vergnaud G, Pourcel C. 2007. Multiple-locus variable-number tandem-repeat analysis for longitudinal survey of sources of *Pseudomonas aeruginosa* infection in cystic fibrosis patients. J Clin Microbiol 45:3175-83.
94. Li W, Raoult D, Fournier PE. 2009. Bacterial strain typing in the genomic era. FEMS Microbiol Rev 33:892-916.
95. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Kohler T, van Delden C, Weinel C, Slickers P, Tummler B. 2007. Population structure of *Pseudomonas aeruginosa*. Proc Natl Acad Sci U S A 104:8101-6.
96. Vernez I, Hauser P, Bernasconi MV, Blanc DS. 2005. Population genetic analysis of *Pseudomonas aeruginosa* using multilocus sequence typing. FEMS Immunol Med Microbiol 43:29-35.
97. Dettman JR, Rodrigue N, Aaron SD, Kassen R. 2013. Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. Proc Natl Acad Sci U S A 110:21065-70.
98. Jelsbak L, Johansen HK, Frost AL, Thogersen R, Thomsen LE, Ciofu O, Yang L, Haagensen JA, Hoiby N, Molin S. 2007. Molecular epidemiology and dynamics of *Pseudomonas aeruginosa* populations in lungs of cystic fibrosis patients. Infect Immun 75:2214-24.

99. Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, Pallen MJ, Ala'Aldeen DA. 2013. Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. Euro Surveill 18.
100. Erbay H, Yalcin AN, Serin S, Turgut H, Tomatir E, Cetin B, Zencir M. 2003. Nosocomial infections in intensive care unit in a Turkish university hospital: a 2-year survey. Intensive Care Med 29:1482-8.
101. Gaynes R, Edwards JR, National Nosocomial Infections Surveillance S. 2005. Overview of nosocomial infections caused by gram-negative bacilli. Clin Infect Dis 41:848-54.
102. Berthelot P, Grattard F, Mahul P, Pain P, Jospe R, Venet C, Carricajo A, Aubert G, Ros A, Dumont A, Lucht F, Zeni F, Auboyer C, Bertrand JC, Pozzetto B. 2001. Prospective study of nosocomial colonization and infection due to *Pseudomonas aeruginosa* in mechanically ventilated patients. Intensive Care Med 27:503-12.
103. Lari AR, Alaghehbandan R. 2000. Nosocomial infections in an Iranian burn care center. Burns 26:737-40.
104. Yildirim S, Nursal TZ, Tarim A, Torer N, Noyan T, Demiroglu YZ, Moray G, Haberal M. 2005. Bacteriological profile and antibiotic resistance: comparison of findings in a burn intensive care unit, other intensive care units, and the hospital services unit of a single center. J Burn Care Rehabil 26:488-92.
105. Bertrand JJ, West JT, Engel JN. 2010. Genetic analysis of the regulation of type IV pilus function by the Chp chemosensory system of *Pseudomonas aeruginosa*. J Bacteriol 192:994-1010.
106. Cobben NA, Drent M, Jonkers M, Wouters EF, Vaneechoutte M, Stobberingh EE. 1996. Outbreak of severe *Pseudomonas aeruginosa* respiratory infections due to contaminated nebulizers. J Hosp Infect 33:63-70.
107. Lanini S, D'Arezzo S, Puro V, Martini L, Imperi F, Piselli P, Montanaro M, Paoletti S, Visca P, Ippolito G. 2011. Molecular epidemiology of a *Pseudomonas aeruginosa* hospital outbreak driven by a contaminated disinfectant-soap dispenser. PLoS One 6:e17064.
108. Morrison AJ, Jr., Wenzel RP. 1984. Epidemiology of infections due to *Pseudomonas aeruginosa*. Rev Infect Dis 6 Suppl 3:S627-42.
109. Blanc DS, Nahimana I, Petignat C, Wenger A, Bille J, Francioli P. 2004. Faucets as a reservoir of endemic *Pseudomonas aeruginosa* colonization/infections in intensive care units. Intensive Care Med 30:1964-8.
110. Petignat C, Francioli P, Nahimana I, Wenger A, Bille J, Schaller MD, Revelly JP, Zanetti G, Blanc DS. 2006. Exogenous sources of pseudomonas aeruginosa in intensive care unit patients: implementation of infection control measures and follow-up with molecular typing. Infect Control Hosp Epidemiol 27:953-7.
111. Cuttelod M, Senn L, Terletskiy V, Nahimana I, Petignat C, Eggimann P, Bille J, Prod'hom G, Zanetti G, Blanc DS. 2011. Molecular epidemiology of *Pseudomonas aeruginosa* in intensive care units over a 10-year period (1998-2007). Clin Microbiol Infect 17:57-62.
112. Tissot F, Blanc DS, Basset P, Zanetti G, Berger MM, Que YA, Eggimann P, Senn L. 2016. New genotyping method discovers sustained nosocomial *Pseudomonas aeruginosa* outbreak in an intensive care burn unit. J Hosp Infect 94:2-7.
113. Gould D, Chamberlaine A. 1995. Staphylococcus aureus: a review of the literature. J Clin Nurs 4:5-12.

114. van Belkum A, Verkaik NJ, de Vogel CP, Boelens HA, Verveer J, Nouwen JL, Verbrugh HA, Wertheim HF. 2009. Reclassification of *Staphylococcus aureus* nasal carriage types. J Infect Dis 199:1820-6.
115. Williams RE, Jevons MP, Shooter RA, Hunter CJ, Girling JA, Griffiths JD, Taylor GW. 1959. Nasal staphylococci and sepsis in hospital patients. Br Med J 2:658-62.
116. Lakhundi S, Zhang K. 2018. Methicillin-Resistant *Staphylococcus aureus*: Molecular Characterization, Evolution, and Epidemiology. Clin Microbiol Rev 31.
117. Seybold U, Kourbatova EV, Johnson JG, Halvosa SJ, Wang YF, King MD, Ray SM, Blumberg HM. 2006. Emergence of community-associated methicillin-resistant *Staphylococcus aureus* USA300 genotype as a major cause of health care-associated blood stream infections. Clin Infect Dis 42:647-56.
118. Peton V, Le Loir Y. 2014. *Staphylococcus aureus* in veterinary medicine. Infect Genet Evol 21:602-15.
119. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, Moore CE, Day NP. 2003. How clonal is Staphylococcus aureus? J Bacteriol 185:3307-16.
120. Feil EJ, Smith JM, Enright MC, Spratt BG. 2000. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics 154:1439-50.
121. Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann E, Dyrek I, Achtman M. 1998. Free recombination within *Helicobacter pylori.* Proc Natl Acad Sci U S A 95:12619-24.
122. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama A, Mizutani-Ui Y, Takahashi NK, Sawano T, Inoue R, Kaito C, Sekimizu K, Hirakawa H, Kuhara S, Goto S, Yabuzaki J, Kanehisa M, Yamashita A, Oshima K, Furuya K, Yoshino C, Shiba T, Hattori M, Ogasawara N, Hayashi H, Hiramatsu K. 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357:1225-40.
123. Baba T, Takeuchi F, Kuroda M, Yuzawa H, Aoki K, Oguchi A, Nagai Y, Iwama N, Asano K, Naimi T, Kuroda H, Cui L, Yamamoto K, Hiramatsu K. 2002. Genome and virulence determinants of high virulence community-acquired MRSA. Lancet 359:1819-27.
124. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, Dodson RJ, Daugherty SC, Madupu R, Angiuoli SV, Durkin AS, Haft DH, Vamathevan J, Khouri H, Utterback T, Lee C, Dimitrov G, Jiang L, Qin H, Weidman J, Tran K, Kang K, Hance IR, Nelson KE, Fraser CM. 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J Bacteriol 187:2426-38.
125. Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R, Barron A, Bason N, Bentley SD, Chillingworth C, Chillingworth T, Churcher C, Clark L, Corton C, Cronin A, Doggett J, Dowd L, Feltwell T, Hance Z, Harris B, Hauser H, Holroyd S, Jagels K, James KD, Lennard N, Line A, Mayes R, Moule S, Mungall K, Ormond D, Quail MA, Rabbinowitsch E, Rutherford K, Sanders M, Sharp S, Simmonds M, Stevens K, Whitehead S, Barrell BG, Spratt BG, Parkhill J. 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci U S A 101:9786-91.

126. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. 2008. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. J Bacteriol 190:300-10.
127. Novick RP. 2003. Autoinduction and signal transduction in the regulation of staphylococcal virulence. Mol Microbiol 48:1429-49.
128. Kaneko J, Kamio Y. 2004. Bacterial two-component and hetero-heptameric pore-forming cytolytic toxins: structures, pore-forming mechanism, and organization of the genes. Biosci Biotechnol Biochem 68:981-1003.
129. Liu GY. 2009. Molecular pathogenesis of *Staphylococcus aureus* infection. Pediatr Res 65:71R-77R.
130. Drechsel H, Freund S, Nicholson G, Haag H, Jung O, Zahner H, Jung G. 1993. Purification and chemical characterization of staphyloferrin B, a hydrophilic siderophore from staphylococci. Biometals 6:185-92.
131. Foster TJ. 2005. Immune evasion by staphylococci. Nat Rev Microbiol 3:948-58.
132. von Eiff C, Peters G, Becker K. 2006. The small colony variant (SCV) concept -- the role of staphylococcal SCVs in persistent infections. Injury 37 Suppl 2:S26-33.
133. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, Craig AS, Zell ER, Fosheim GE, McDougal LK, Carey RB, Fridkin SK, Active Bacterial Core surveillance MI. 2007. Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. JAMA 298:1763-71.
134. Klevens RM, Edwards JR, Tenover FC, McDonald LC, Horan T, Gaynes R, National Nosocomial Infections Surveillance S. 2006. Changes in the epidemiology of methicillin-resistant *Staphylococcus aureus* in intensive care units in US hospitals, 1992-2003. Clin Infect Dis 42:389-91.
135. Meyer W. 1966. [Schema for the differentiation of habitat variants of *Staphylococcus aureus*]. Zentralbl Bakteriol Orig 201:465-81.
136. Kapur V, Sischo WM, Greer RS, Whittam TS, Musser JM. 1995. Molecular population genetic analysis of *Staphylococcus aureus* recovered from cows. J Clin Microbiol 33:376-80.
137. Fitzgerald JR, Holden MT. 2016. Genomics of Natural Populations of *Staphylococcus aureus*. Annu Rev Microbiol 70:459-78.
138. Shopsin B, Gomez M, Montgomery SO, Smith DH, Waddington M, Dodge DE, Bost DA, Riehman M, Naidich S, Kreiswirth BN. 1999. Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. J Clin Microbiol 37:3556-63.
139. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Medicine 6:90.
140. Lee DG, Urbach JM, Wu G, Liberati NT, Feinbaum RL, Miyata S, Diggins LT, He J, Saucier M, Deziel E, Friedman L, Li L, Grills G, Montgomery K, Kucherlapati R, Rahme LG, Ausubel FM. 2006. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome Biol 7:R90.
141. Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP. 2016. Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*. J Clin Microbiol 54:333-42.
142. Kanamori H, Parobek CM, Juliano JJ, van Duin D, Cairns BA, Weber DJ, Rutala WA. 2017. A Prolonged Outbreak of KPC-3-Producing *Enterobacter cloacae* and *Klebsiella pneumoniae* Driven by Multiple

Mechanisms of Resistance Transmission at a Large Academic Burn Center. Antimicrob Agents Chemother 61.

143. Parcell BJ, Oravcova K, Pinheiro M, Holden MTG, Phillips G, Turton JF, Gillespie SH. 2018. *Pseudomonas aeruginosa* intensive care unit outbreak: winnowing of transmissions with molecular and genomic typing. J Hosp Infect 98:282-288.
144. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16-21.
145. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15.
146. Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221-4.
147. Carcer DA, Denman SE, McSweeney C, Morrison M. 2011. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. Appl Environ Microbiol 77:8795-8.
148. Carrico JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. 2018. A primer on microbial bioinformatics for nonbioinformaticians. Clin Microbiol Infect 24:342-349.
149. Mellouk FZ, Bakour S, Meradji S, Al-Bayssari C, Bentakouk MC, Zouyed F, Djahoudi A, Boutefnouchet N, Rolain JM. 2017. First Detection of VIM-4-Producing *Pseudomonas aeruginosa* and OXA-48-Producing *Klebsiella pneumoniae* in Northeastern (Annaba, Skikda) Algeria. Microb Drug Resist 23:335-344.
150. Oliver A, Mulet X, Lopez-Causape C, Juan C. 2015. The increasing threat of *Pseudomonas aeruginosa* high-risk clones. Drug Resist Updat 21-22:41-59.
151. Saltykova A, Wuyts V, Mattheus W, Bertrand S, Roosens NHC, Marchal K, De Keersmaecker SCJ. 2018. Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to *Salmonella enterica* serotype Typhimurium and serotype 1,4,[5],12:i. PLoS One 13:e0192504.
152. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML, Zook JM. 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6:235.
153. Kanamori H, Weber DJ, Rutala WA. 2016. Healthcare Outbreaks Associated With a Water Reservoir and Infection Prevention Strategies. Clin Infect Dis 62:1423-35.
154. Kuck P, Mayer C, Wagele JW, Misof B. 2012. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. PLoS One 7:e36593.
155. Jolivet-Gougeon A, Kovacs B, Le Gall-David S, Le Bars H, Bousarghin L, Bonnaure-Mallet M, Lobel B, Guille F, Soussy CJ, Tenke P. 2011. Bacterial hypermutation: clinical implications. J Med Microbiol 60:563-73.
156. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli.* J Clin Microbiol 52:1501-10.

157. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563-9.
158. Anonymous. Amos (Software), The International Encyclopedia of Communication Research Methods doi:doi:10.1002/9781118901731.iecrm0003.
159. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963.
160. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614-24.
161. Hurgobin B, Edwards D. 2017. SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? Biology (Basel) 6.
162. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. 2014. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol 52:2365-70.
163. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA. 2012. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics 13:87.

CHAPTER 7.

Supplementary figures

DLST 1-18 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 20, 10 reads

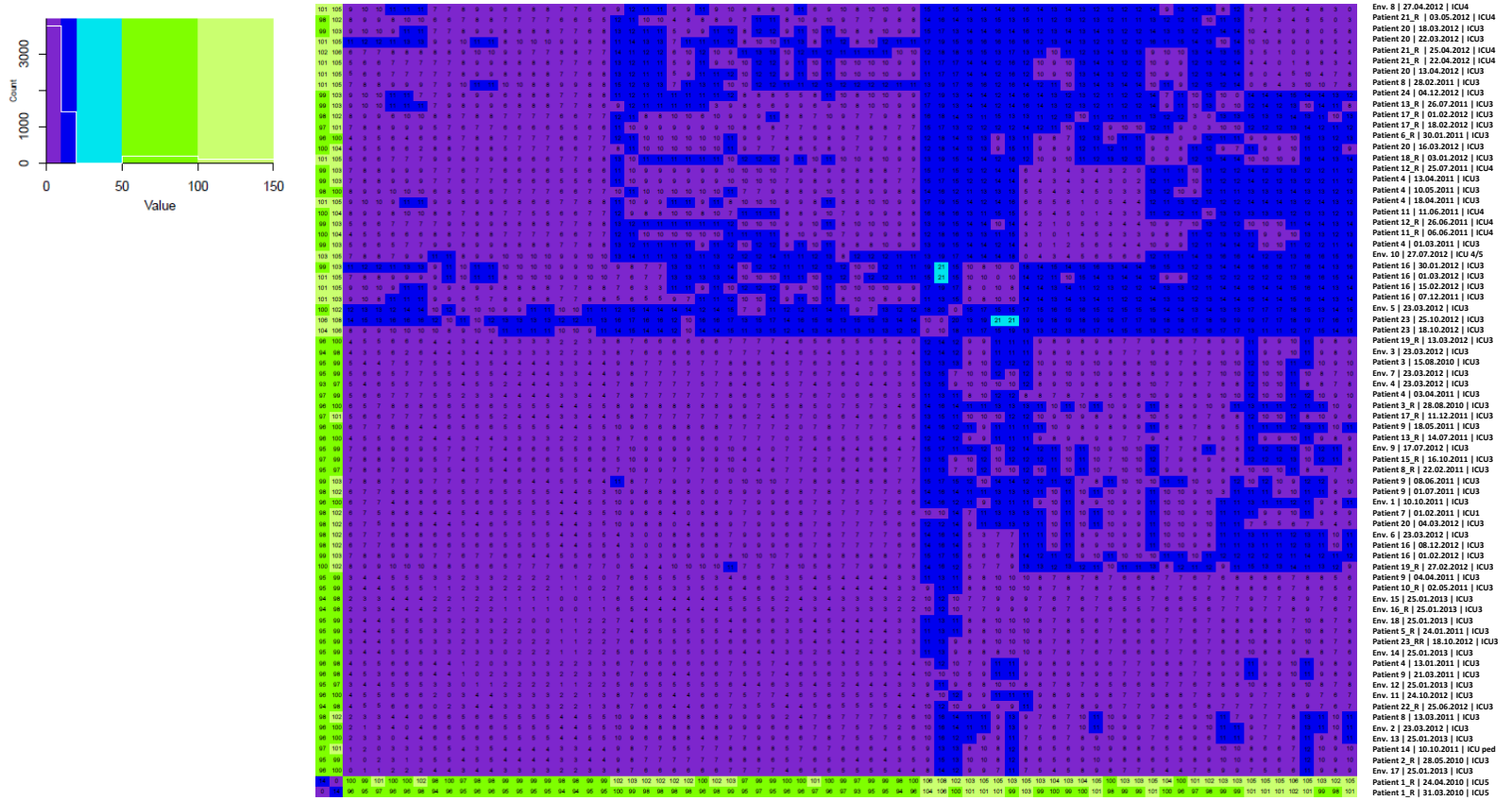


Figure 40. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Env. 8 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-18 phylogeny

Adapted methodology: PA14 reference, MQ 20, 10 reads

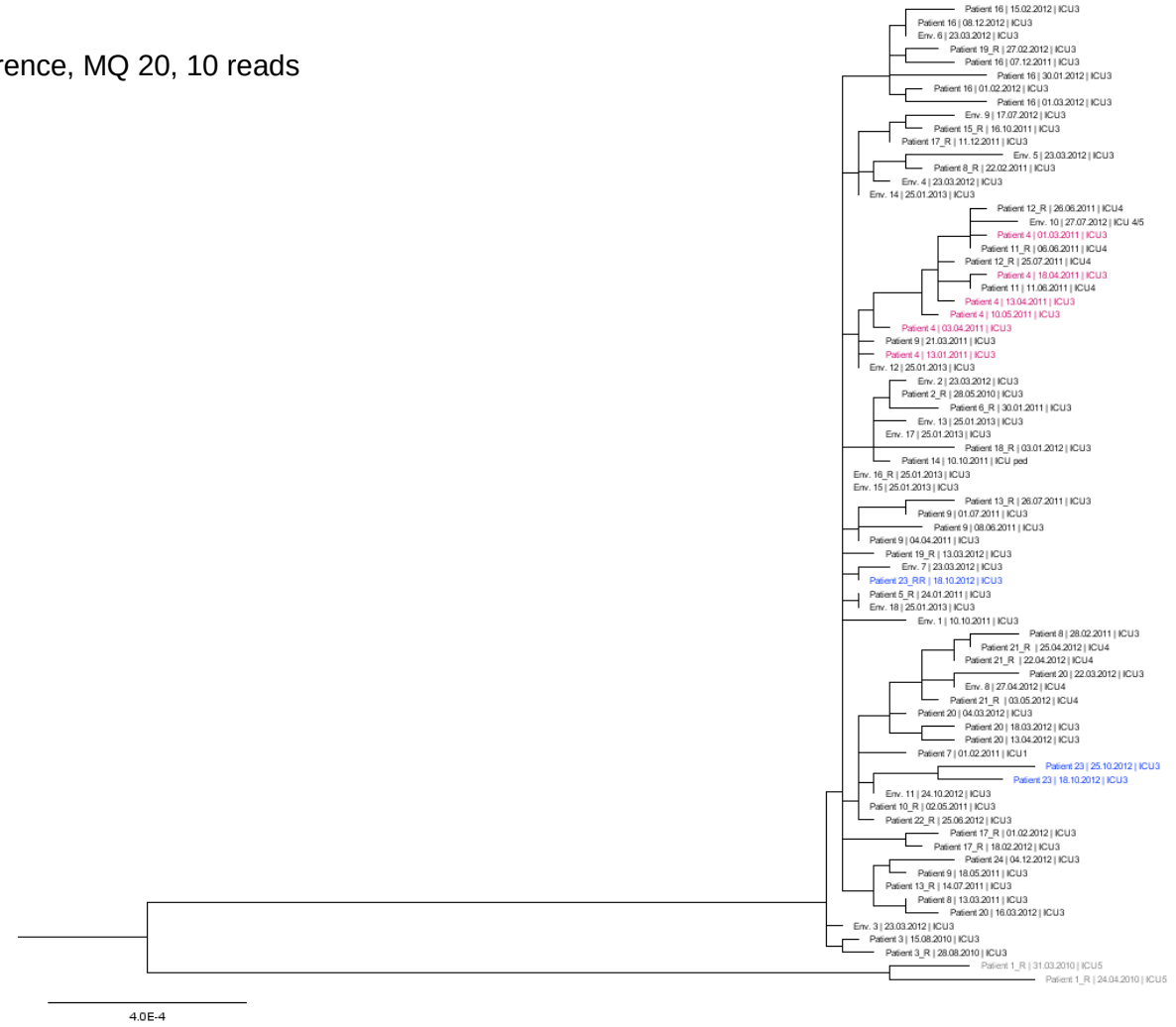


Figure 41. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

DLST 1-21 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 20, 10 reads

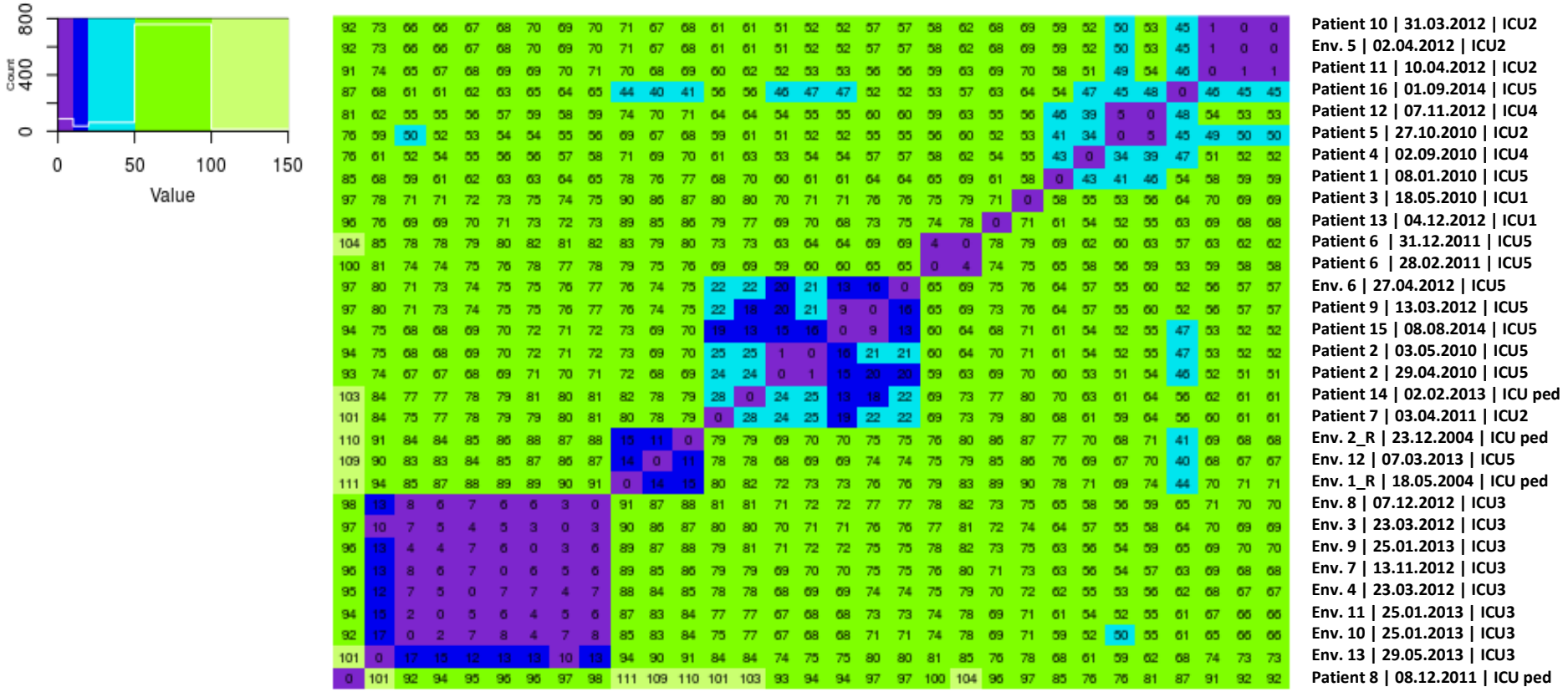


Figure 42. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 10 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PA14 reference, MQ 20, 10 reads

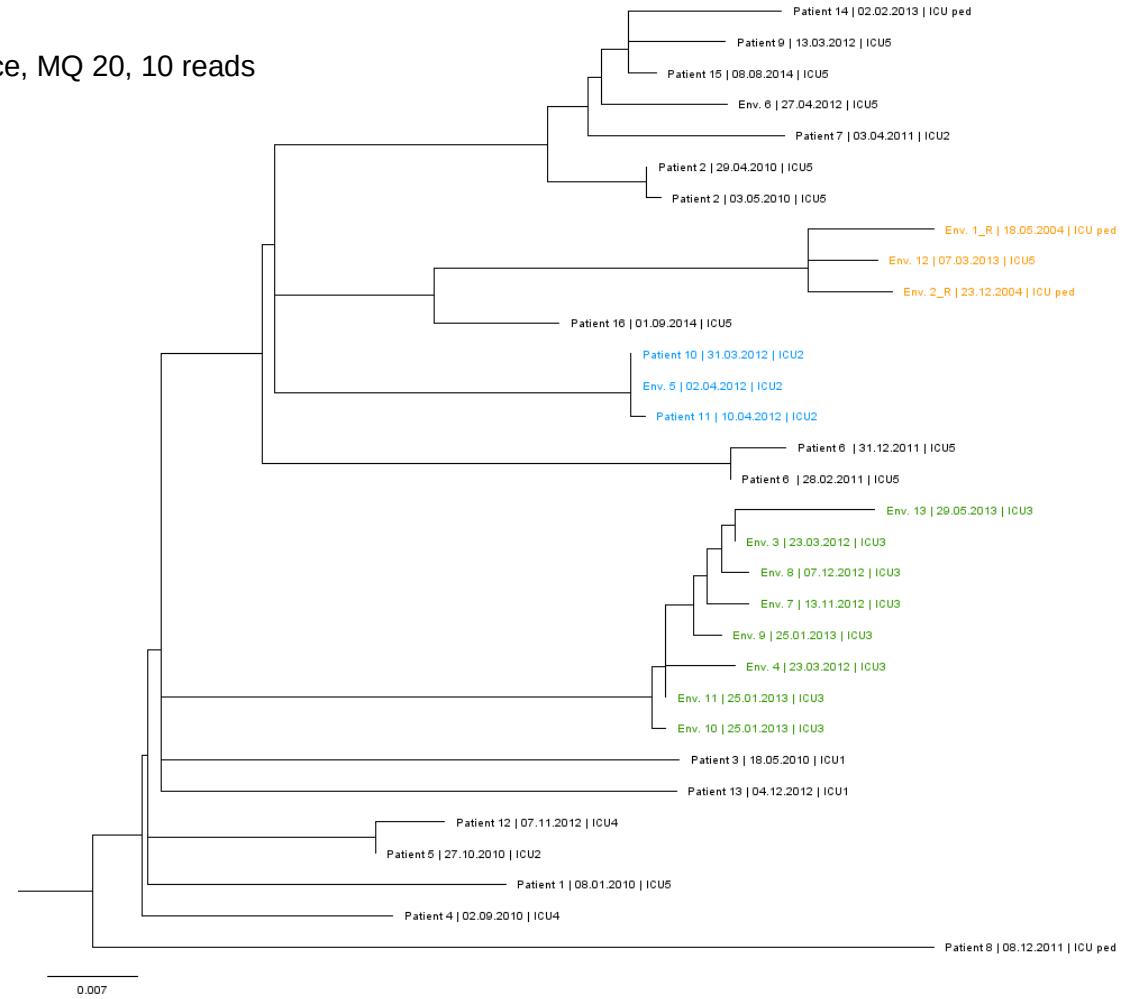


Figure 43. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the *P. aeruginosa* PA14 with a mapping quality of 20 and with a minimum of 10 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 20, 10 reads

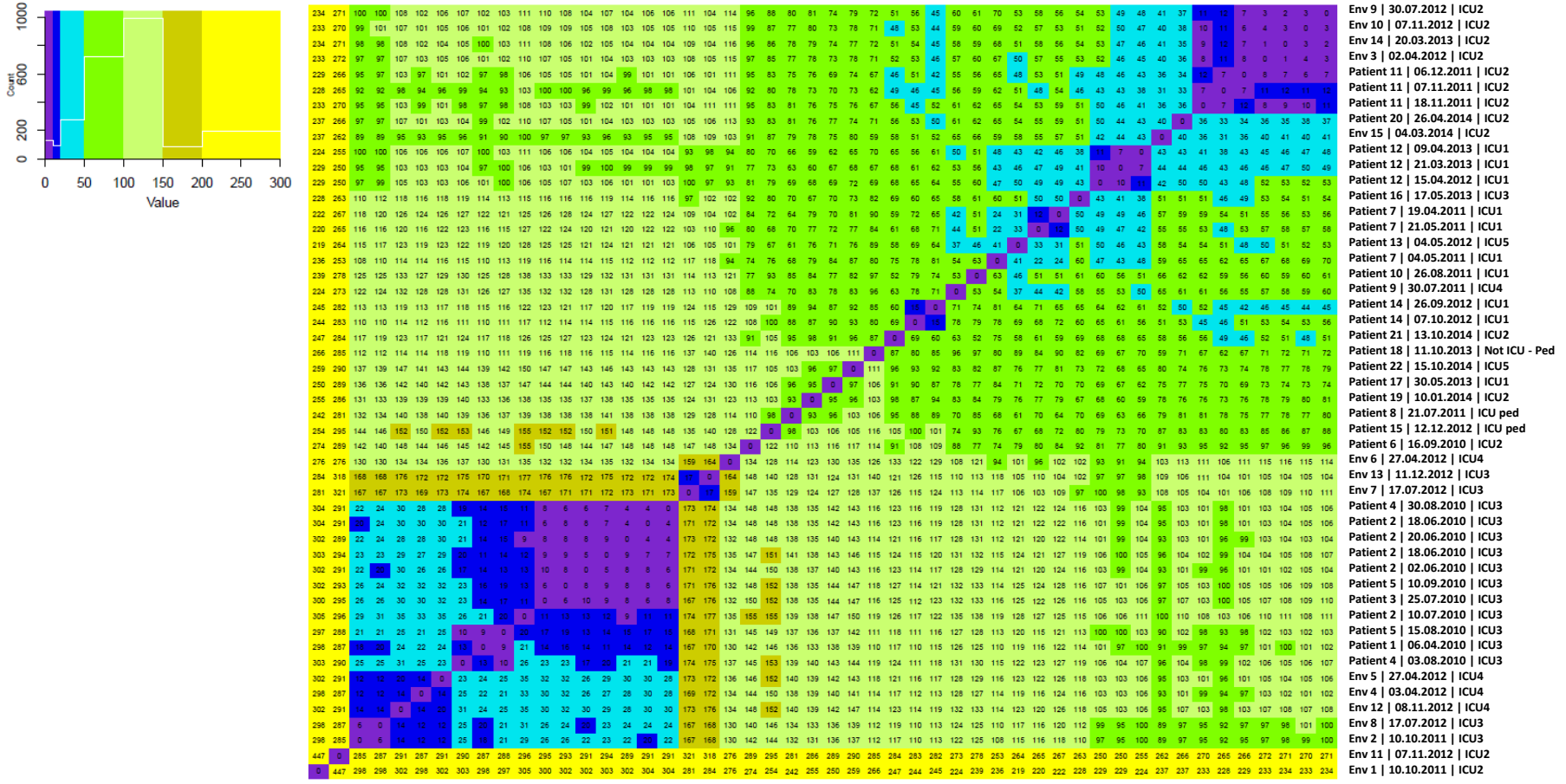


Figure 44. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 10 reads to consider a SNP site.. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Env. 9 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 300. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PA14 reference, MQ 20, 10 reads

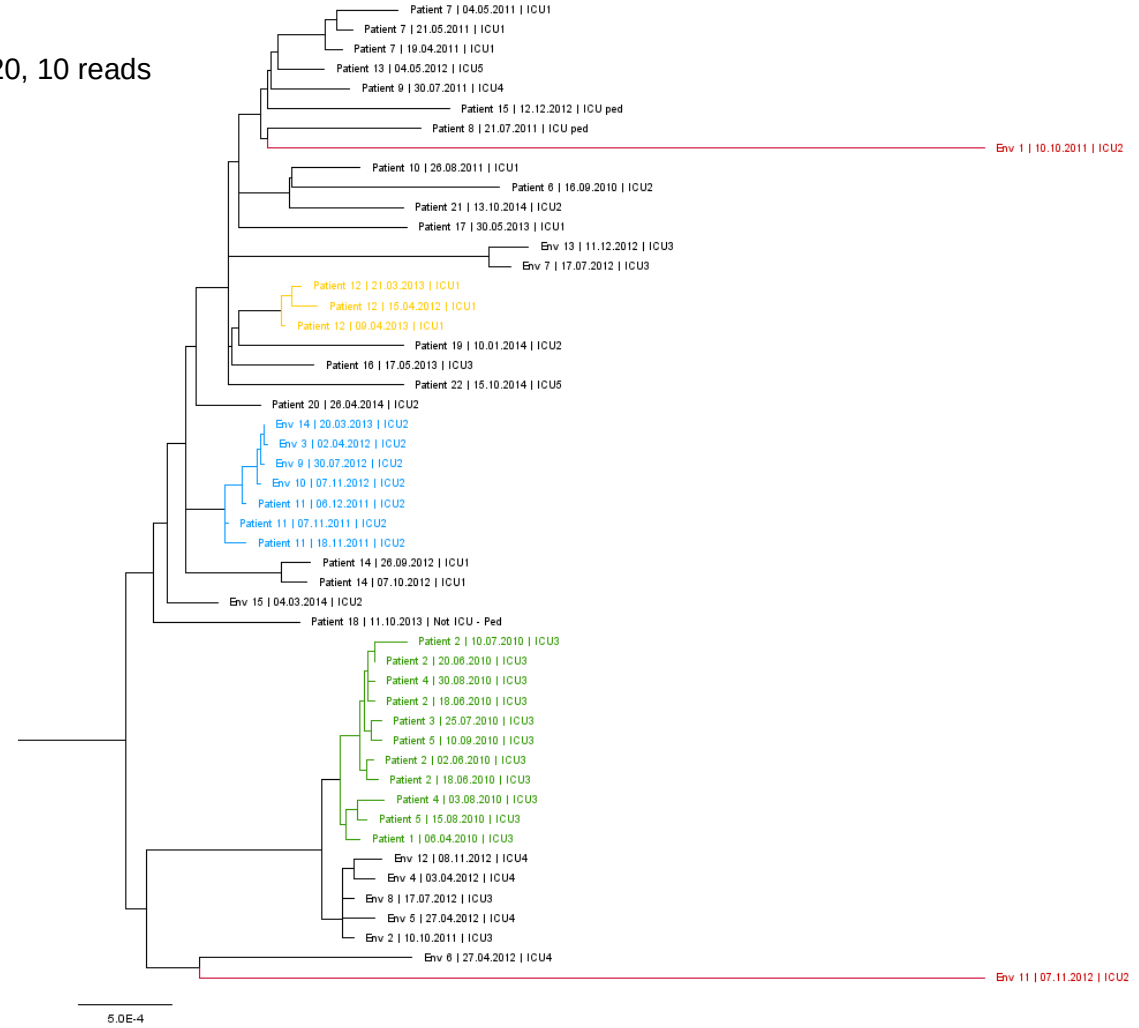


Figure 45. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with a mapping quality of 20 and with a minimum of 10 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red

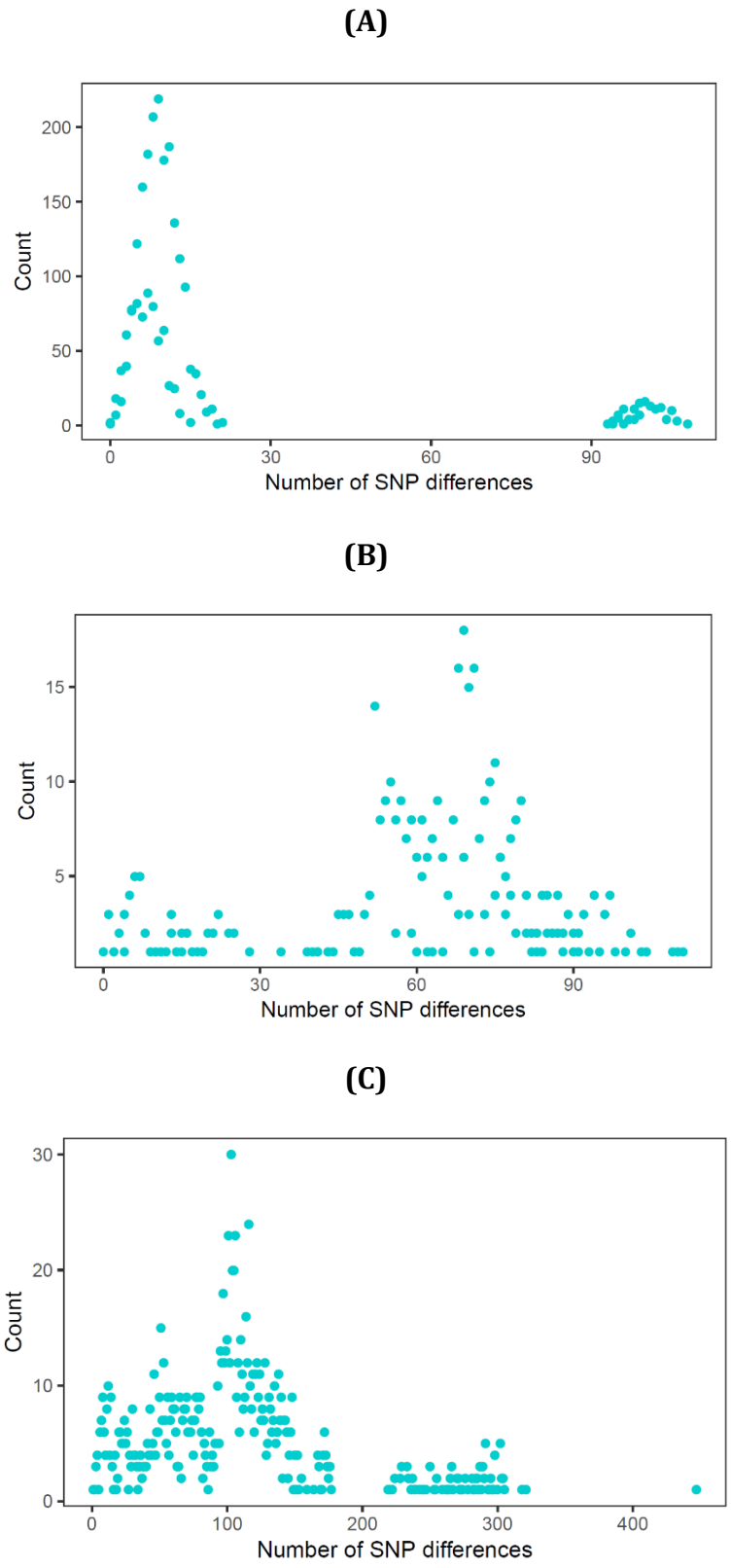


Figure 46. Frequency of number of SNP differences obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 10 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

DLST 1-18 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 20, 20 reads

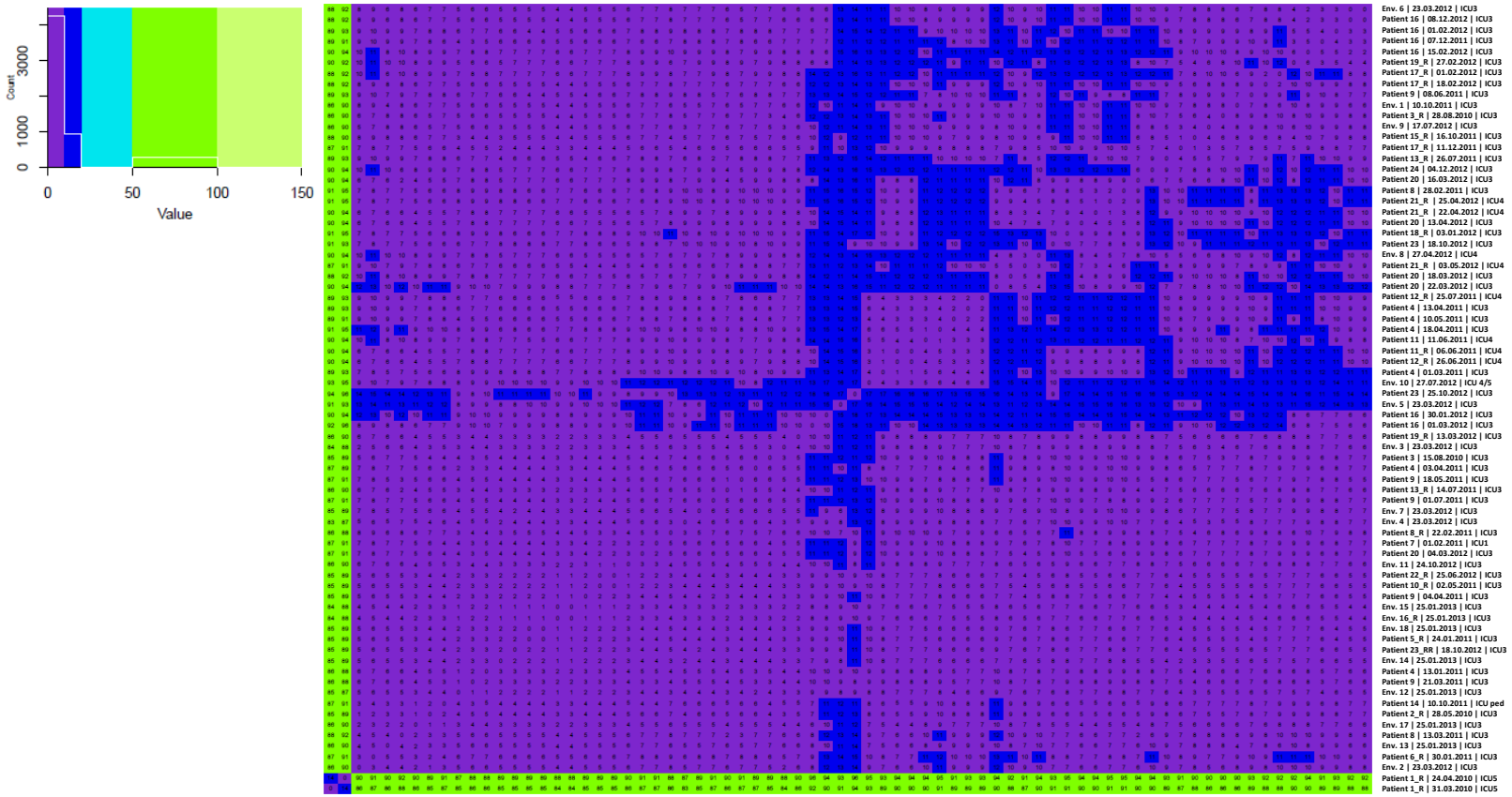


Figure 48. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Env. 6 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-18 phylogeny

Adapted methodology: PA14 reference, MQ 20, 20 reads

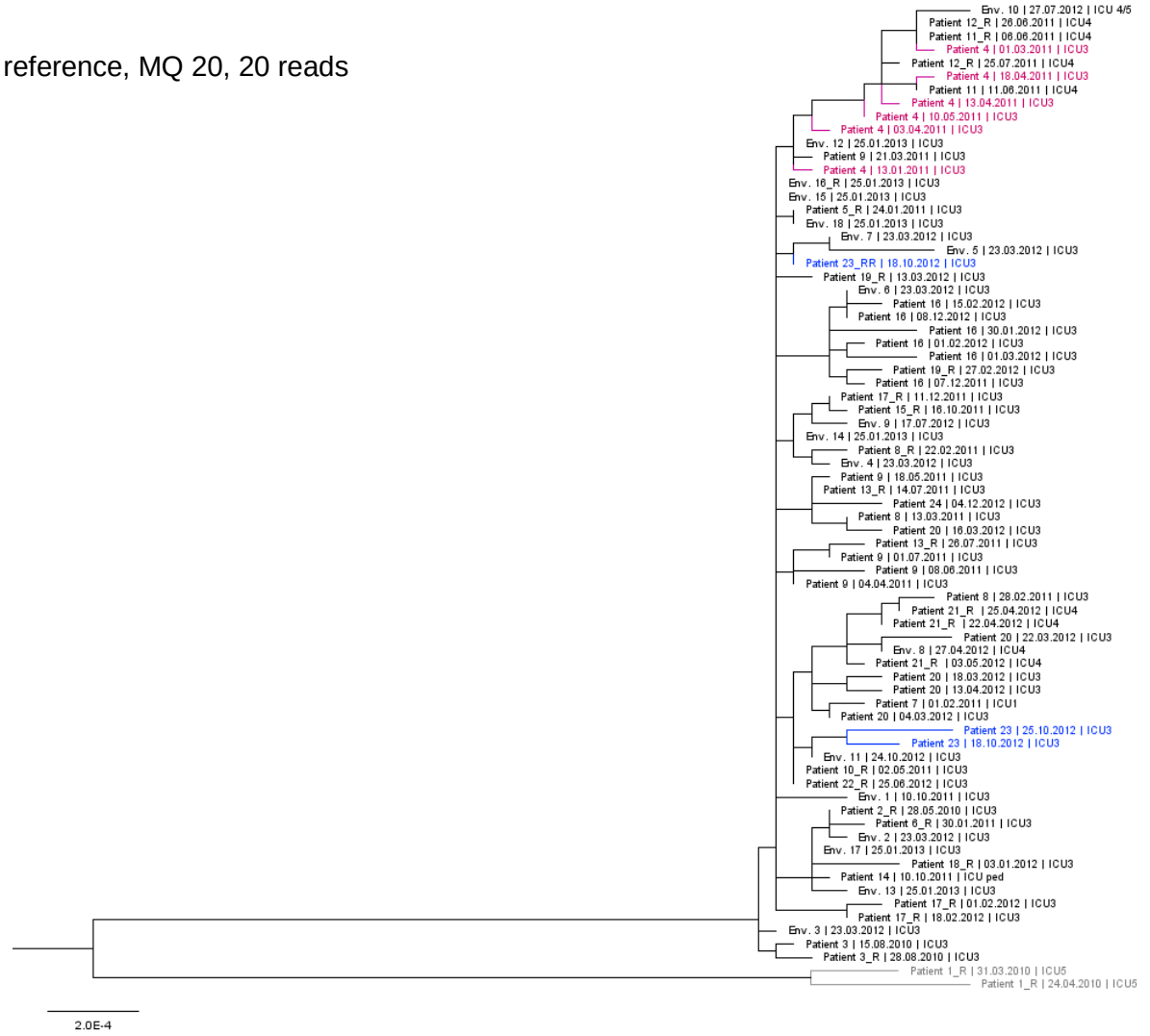


Figure 49. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

DLST 1-21 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 20, 20 reads

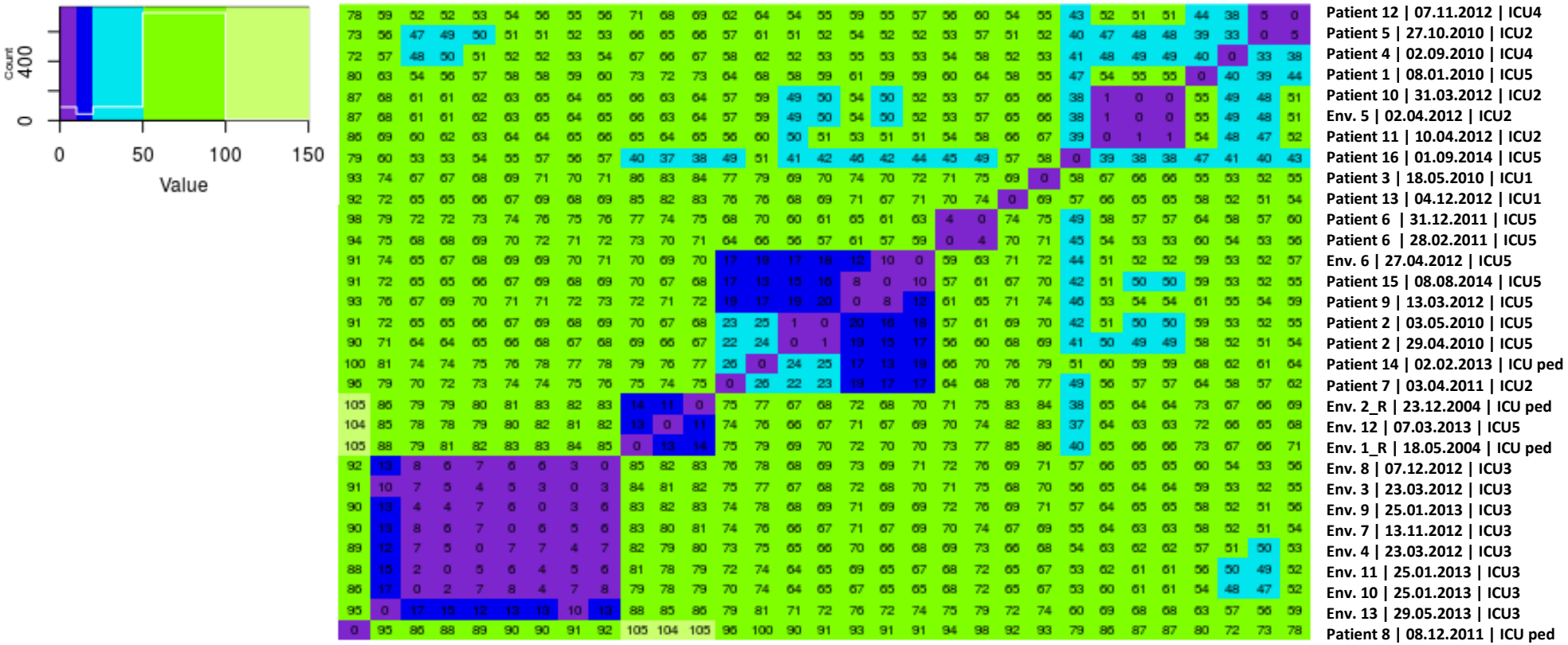


Figure 50. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 12 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PA14 reference, MQ 20, 20reads

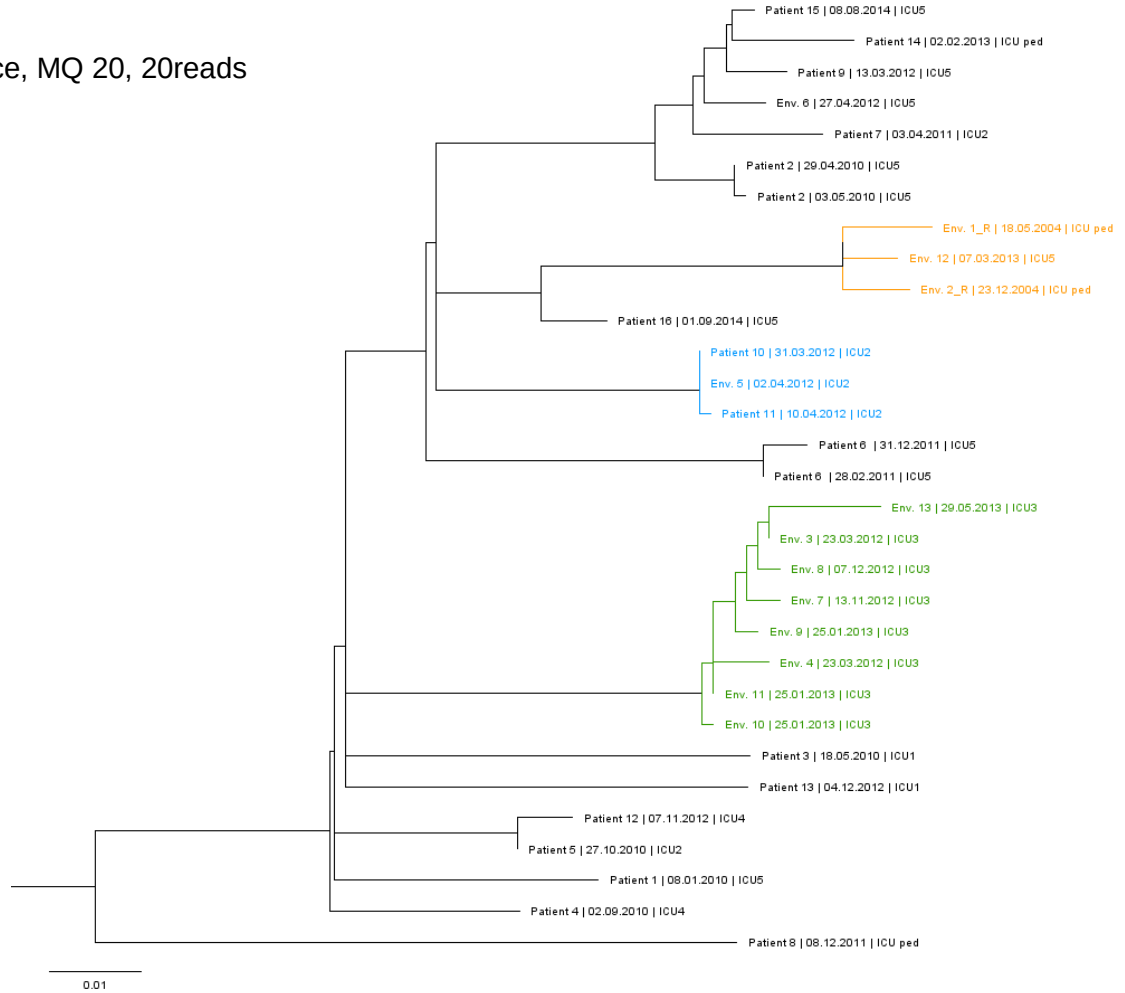


Figure 51. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the *P. aeruginosa* PA14 with a mapping quality of 20 and with a minimum of 20 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 20, 20 reads

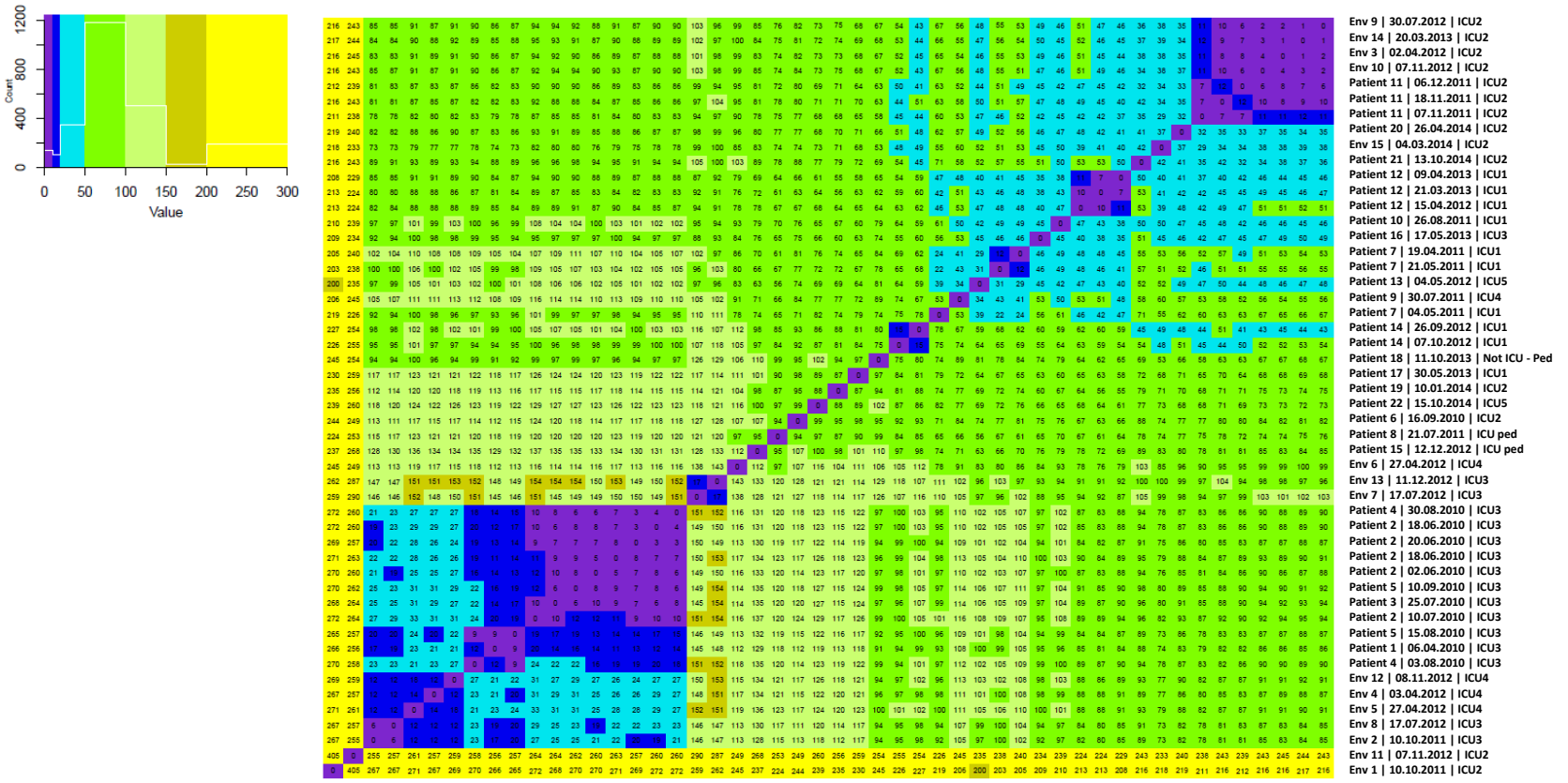


Figure 52. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 20 reads to consider a SNP site.. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Env.9 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 300. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 6-7 phylogeny

Adapted methodology: PA14 reference, MQ 20, 20reads

Figure 53. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with a mapping quality of 20 and with a minimum of 20 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red

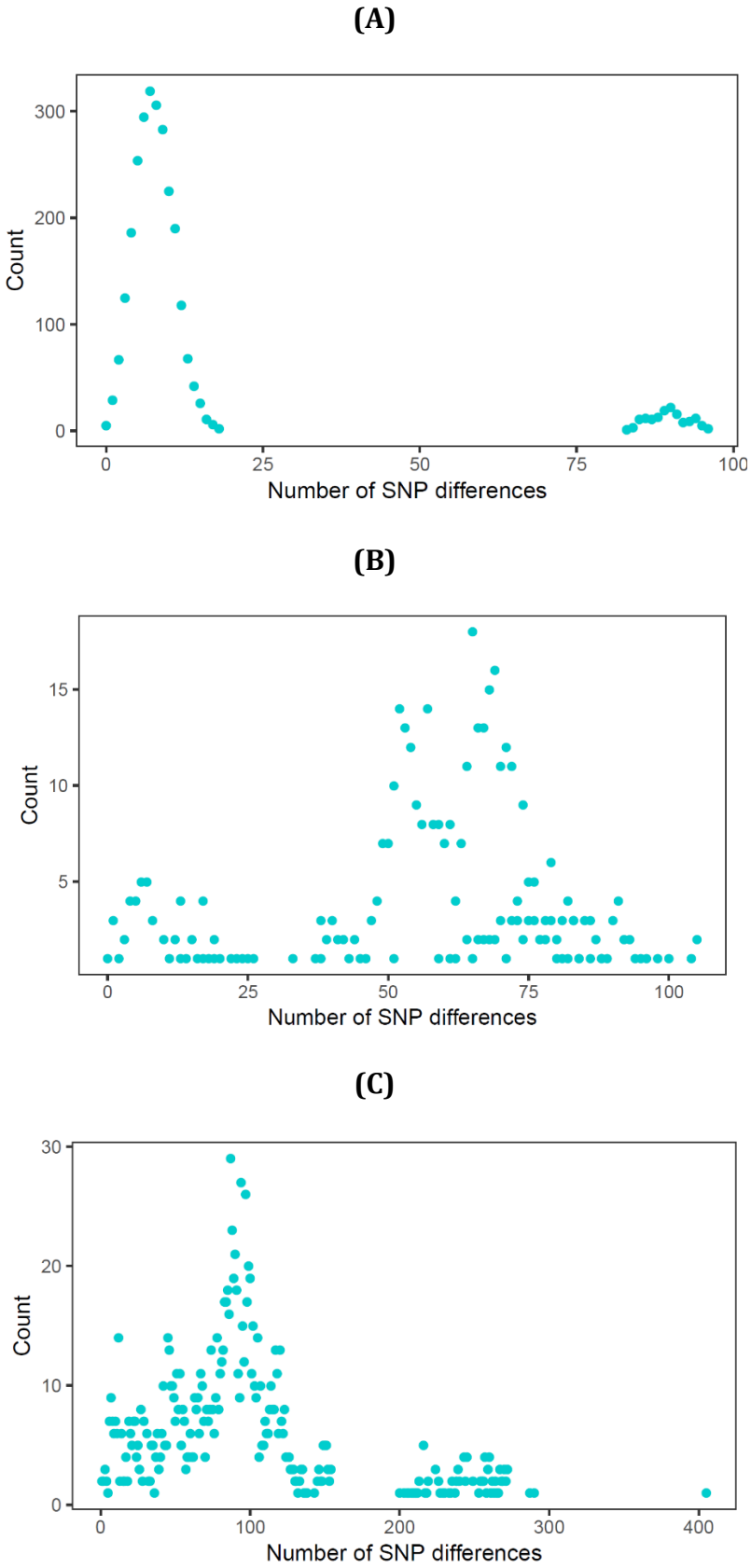


Figure 54. Frequency of number of SNP differences obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 20 and minimum of 20 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

DLST 1-18 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 60, 20 reads

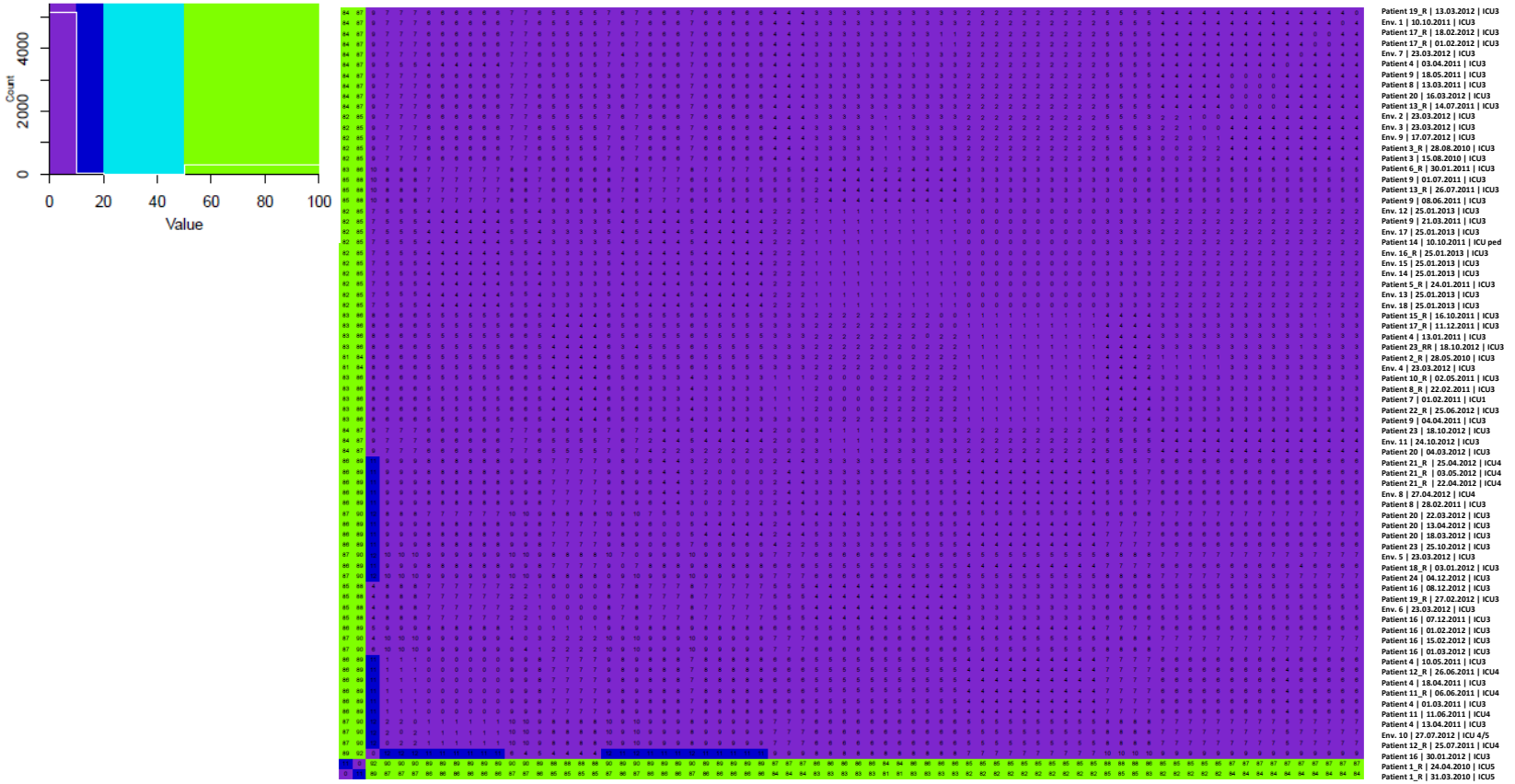


Figure 55. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Patient 19 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, and 100. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-18 phylogeny

Adapted methodology: PA14 reference, MQ 60, 20 reads

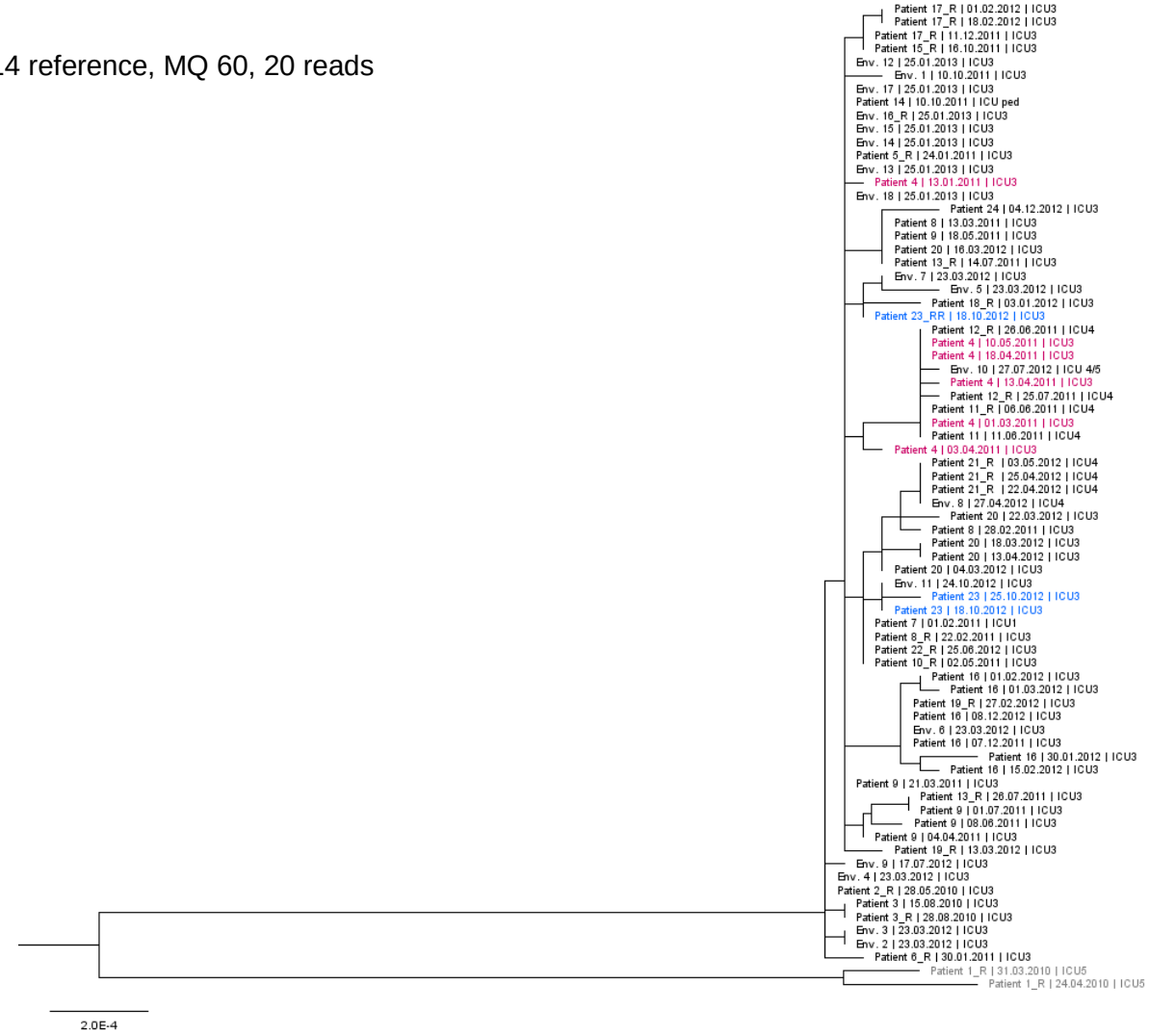


Figure 56. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

DLST 1-21 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 60, 20 reads

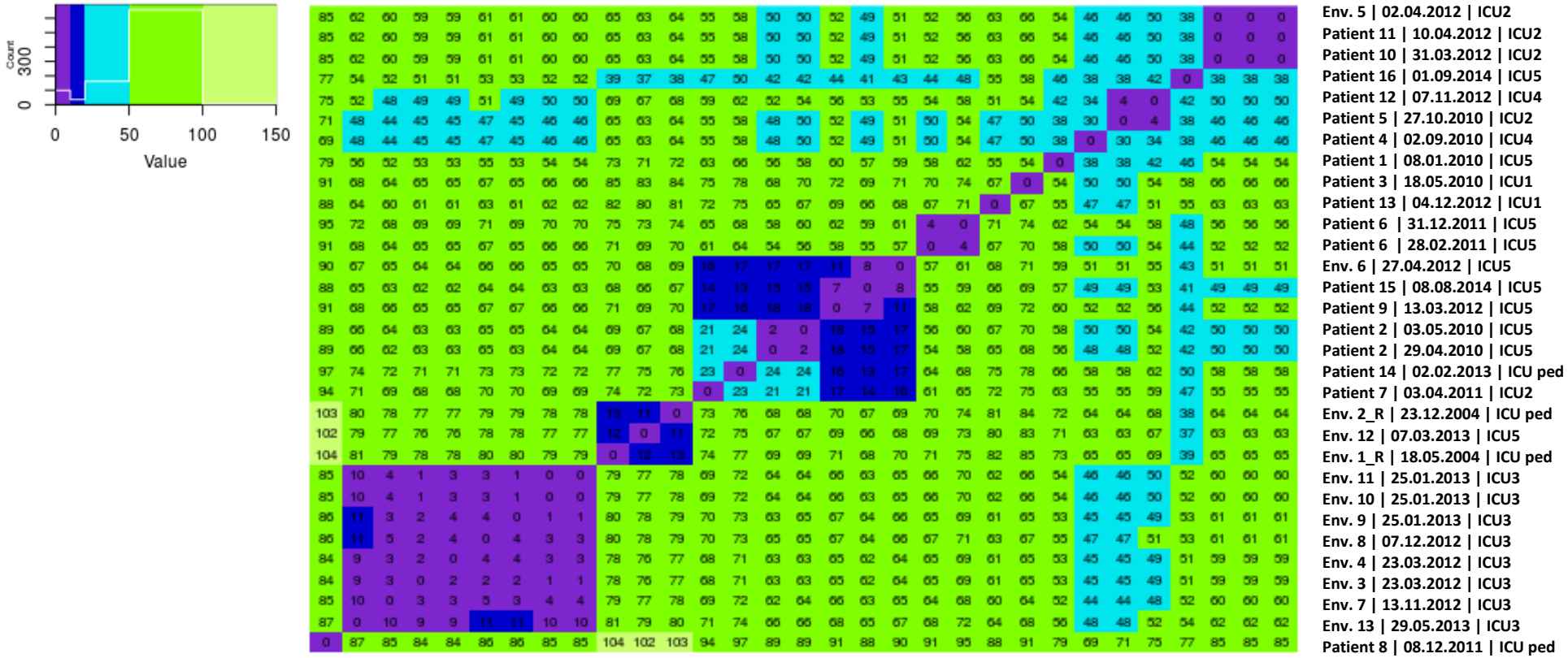


Figure 57. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from: left to right: Patient 8 (first isolate) to Env. 5 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PA14 reference, MQ 60, 20 reads

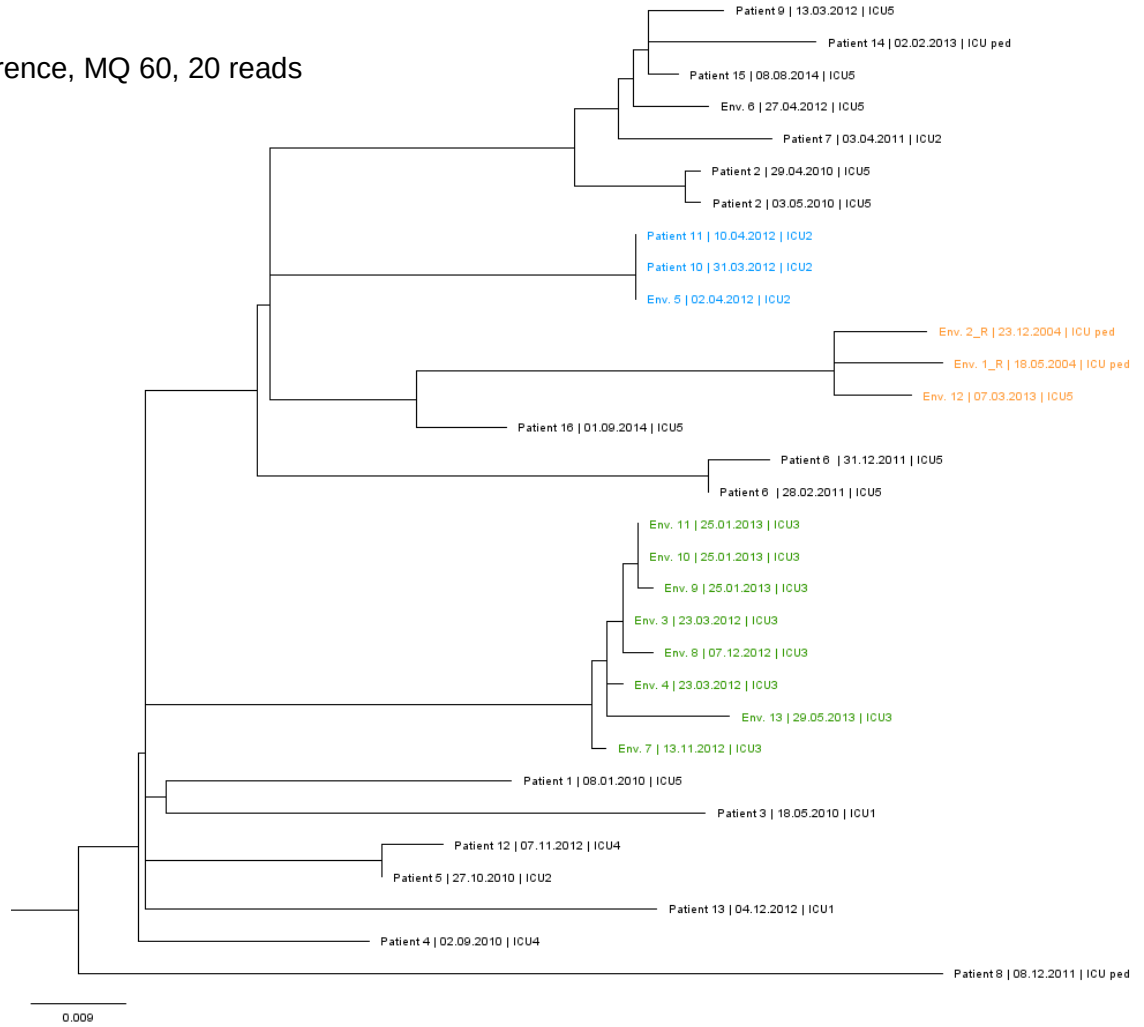


Figure 58. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the *P. aeruginosa* PA14 with a mapping quality of 60 and with a minimum of 20 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 pairwise distance matrix

Adapted methodology: PA14 reference, MQ 60, 20 reads

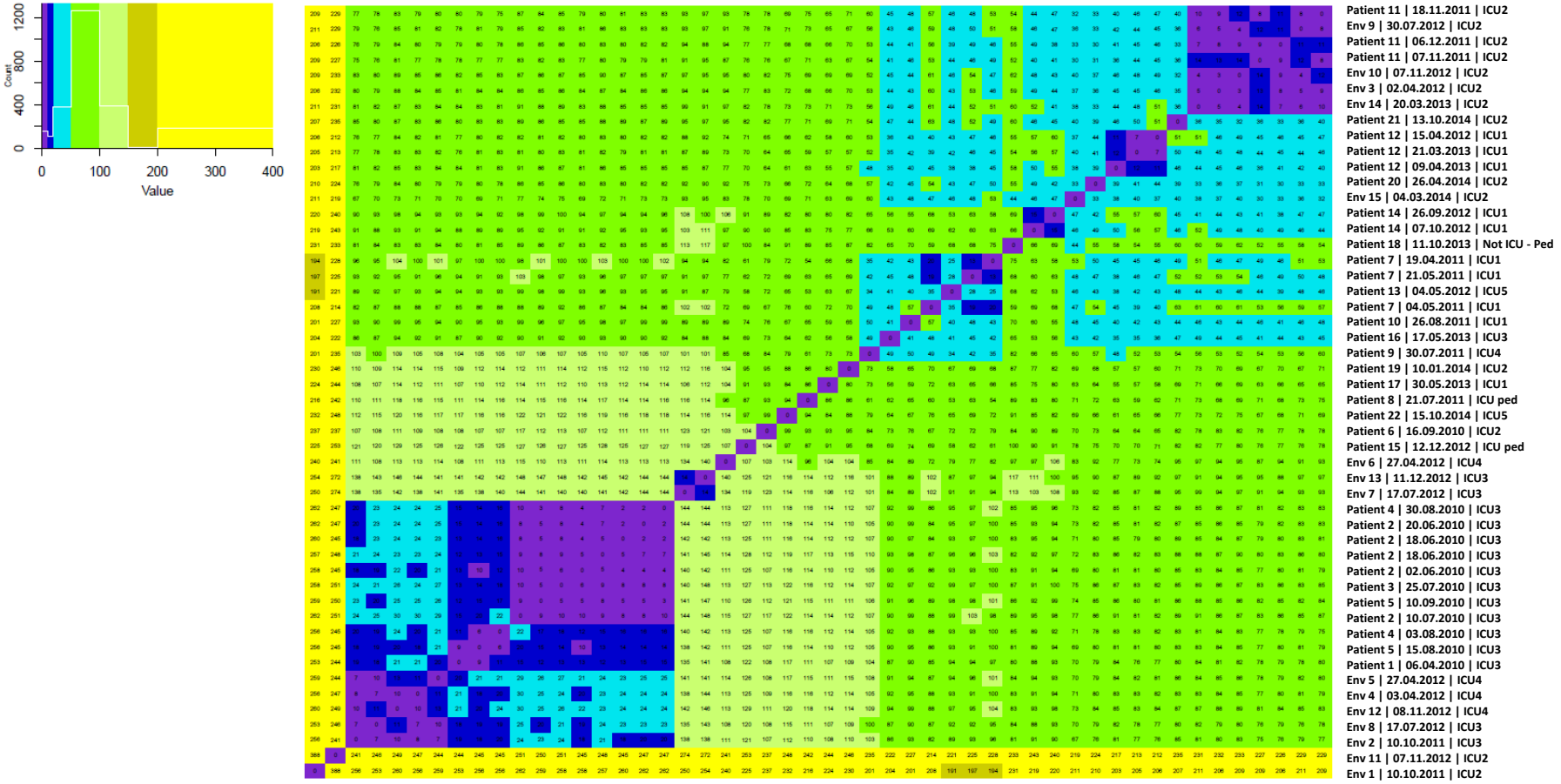


Figure 59. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 400. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 6-7 phylogeny

Adapted methodology: PA14 reference, MQ 60, 20 reads

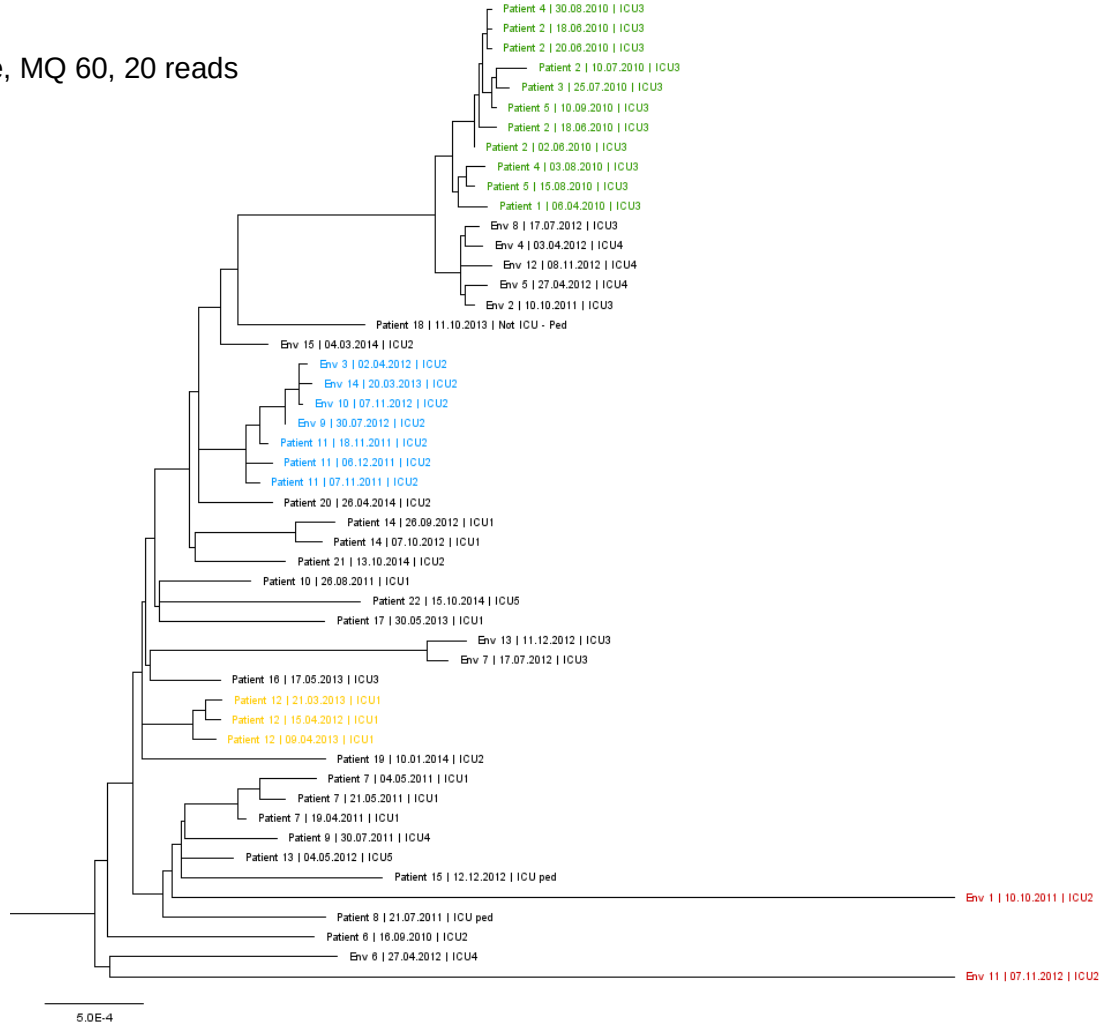
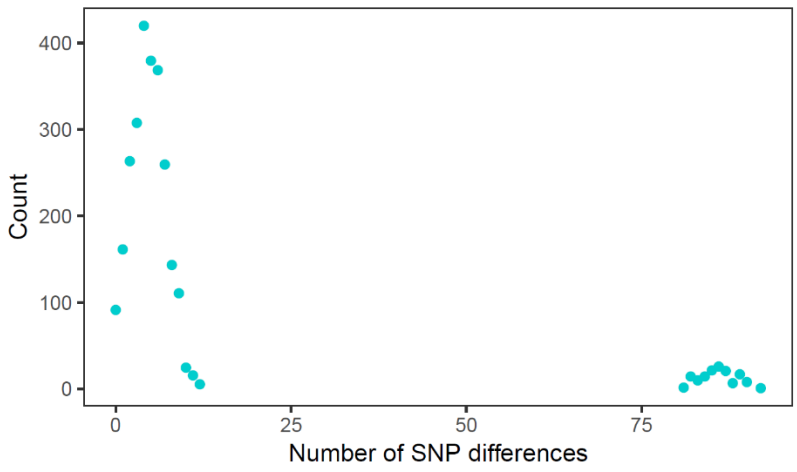
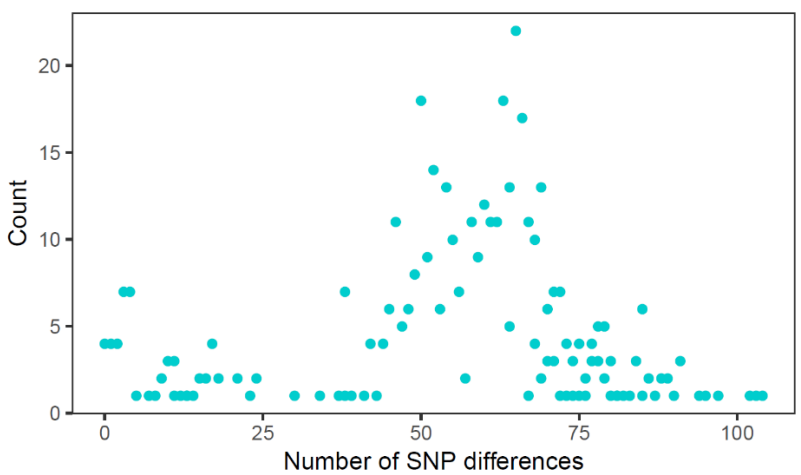


Figure 60. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with a mapping quality of 60 and with a minimum of 20 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red

(A)



(B)



(C)

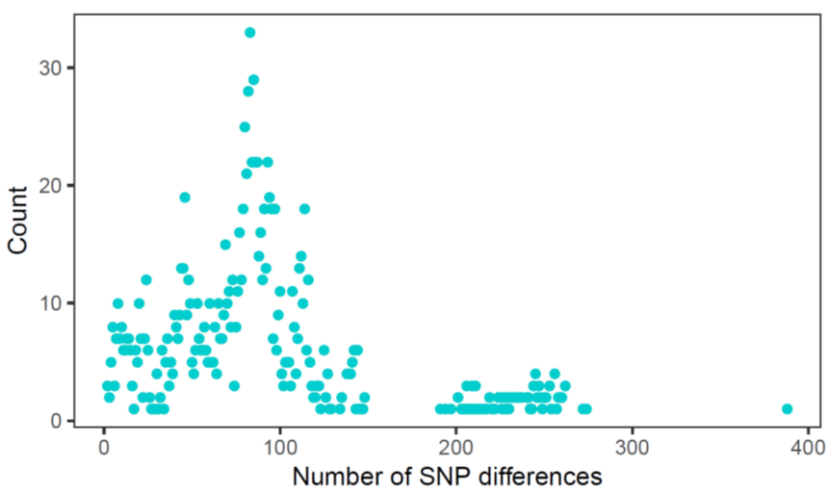


Figure 61. Frequency of number of SNP differences obtained with the adapted methodology, mapping against *P. aeruginosa* PA14 with mapping quality of 60 and minimum of 20 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

DLST 1-18 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 20, 10 reads

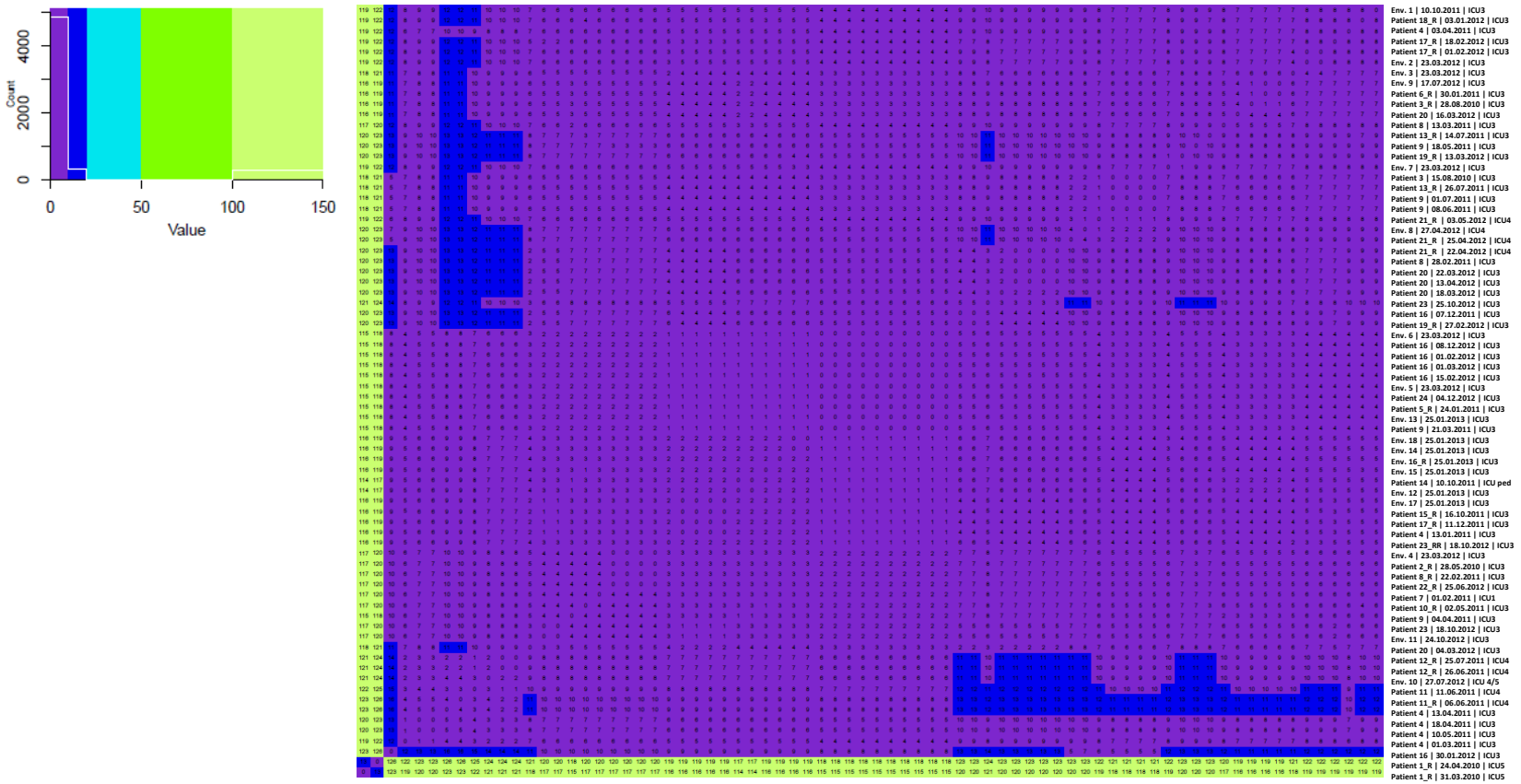


Figure 62. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Env. 1 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-18 phylogeny

Adapted methodology: PacBio reference, MQ 20, 10 reads

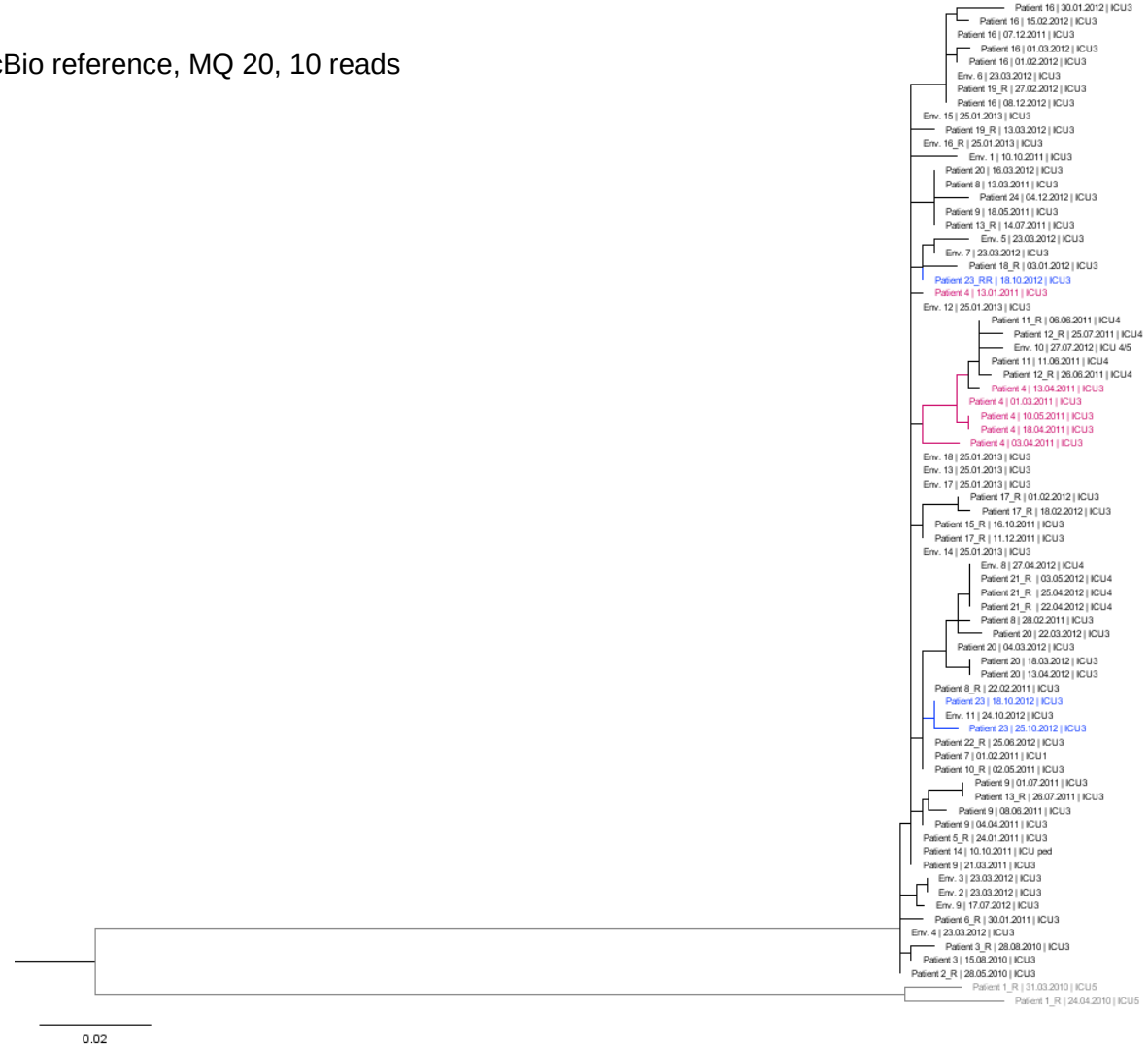


Figure 63. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology mapping against the PacBio reference with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

DLST 1-21 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 20, 10 reads

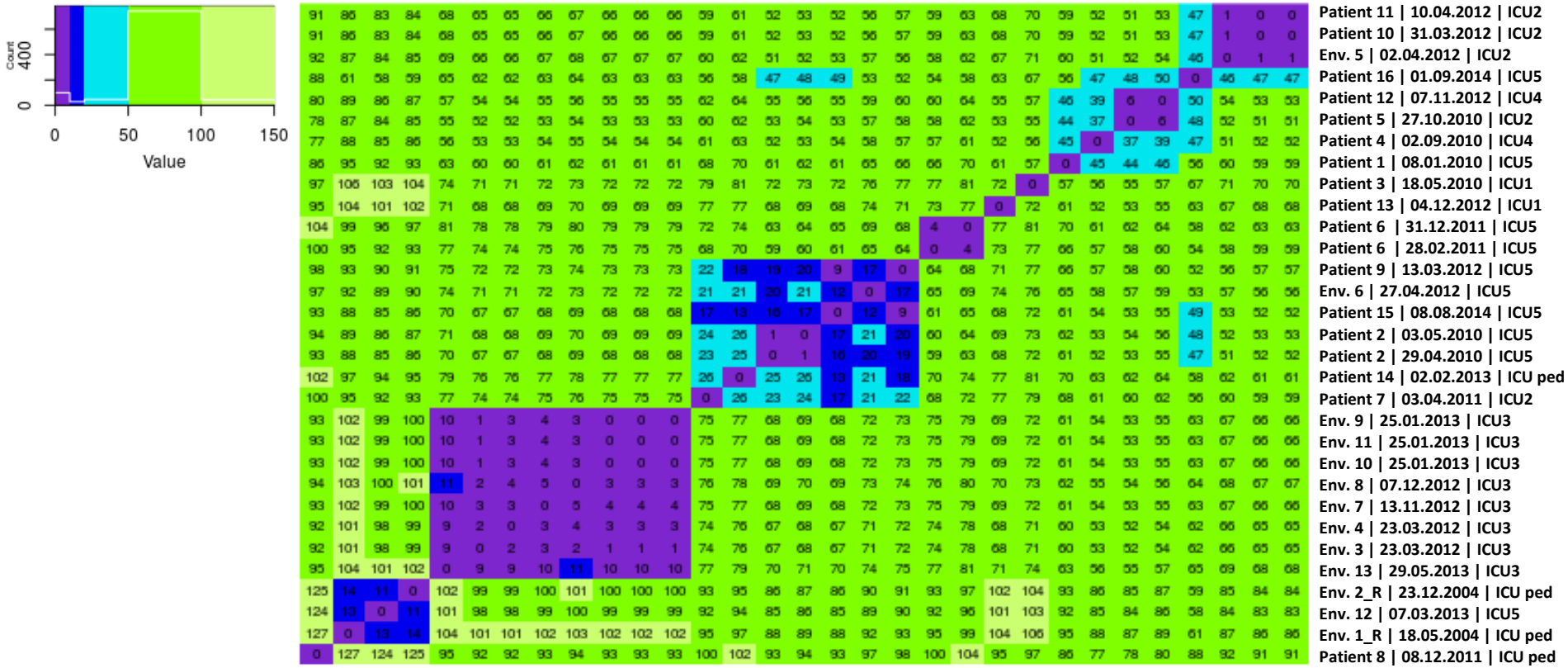


Figure 64. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PacBio reference, MQ 20, 10 reads

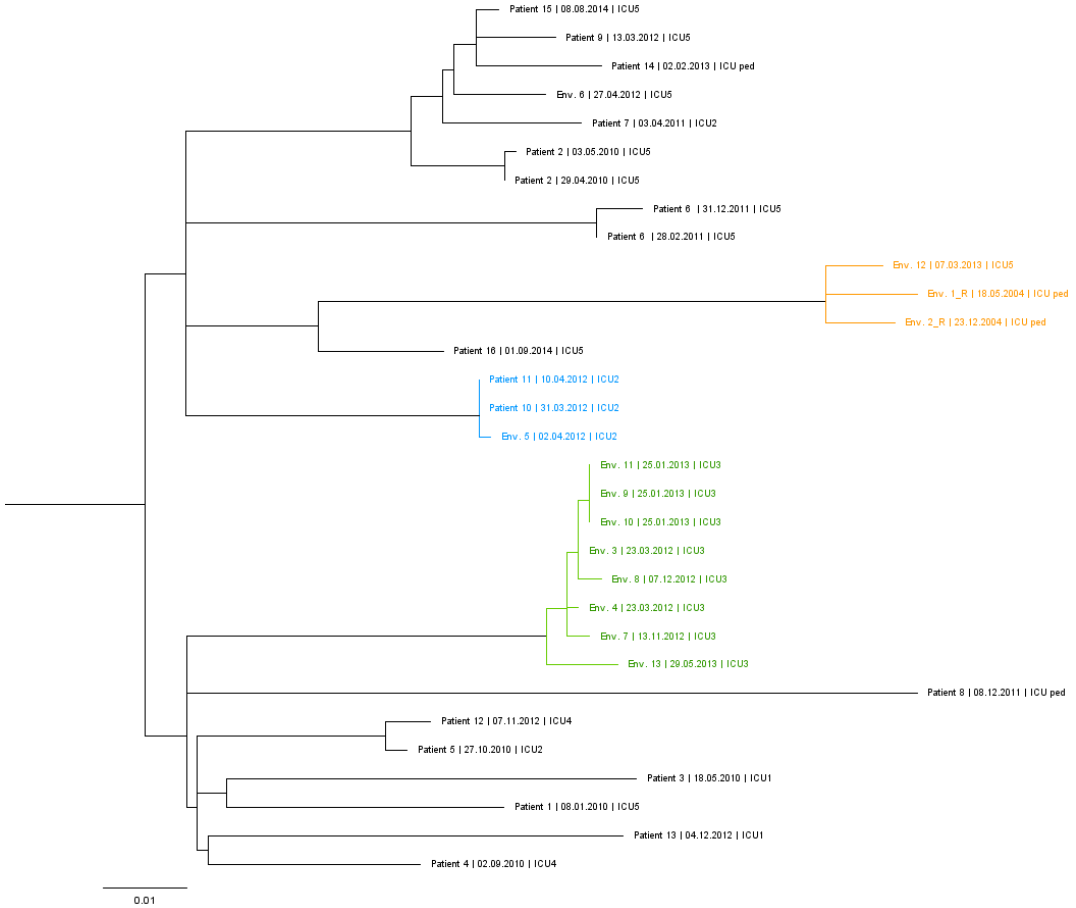


Figure 65. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacBio reference with a mapping quality of 20 and with a minimum of 10 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 20, 10 reads

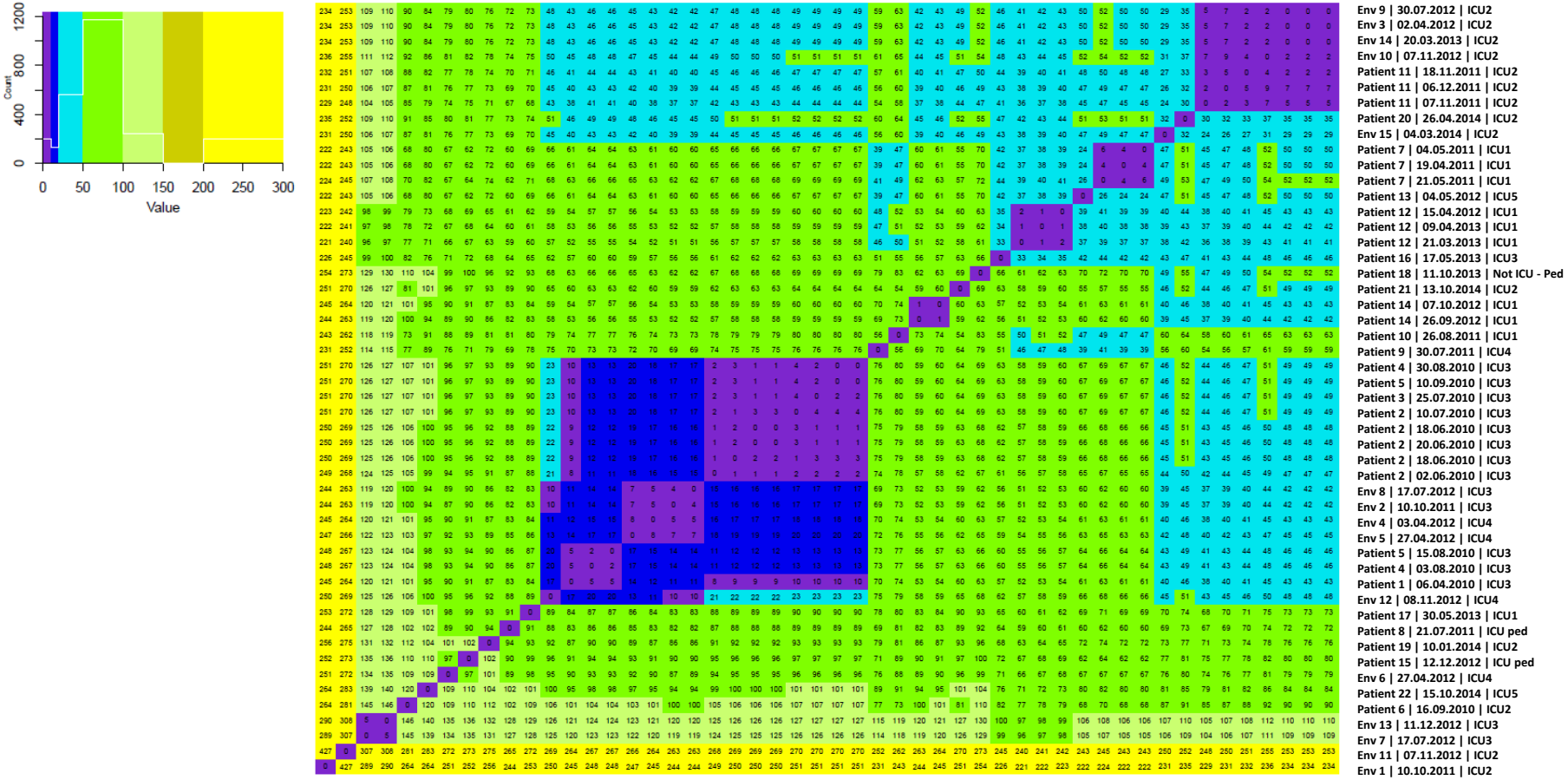


Figure 66. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 10 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Env. 9 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 300. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 6-7 phylogeny

Adapted methodology: PacBio reference, MQ 20, 10 reads

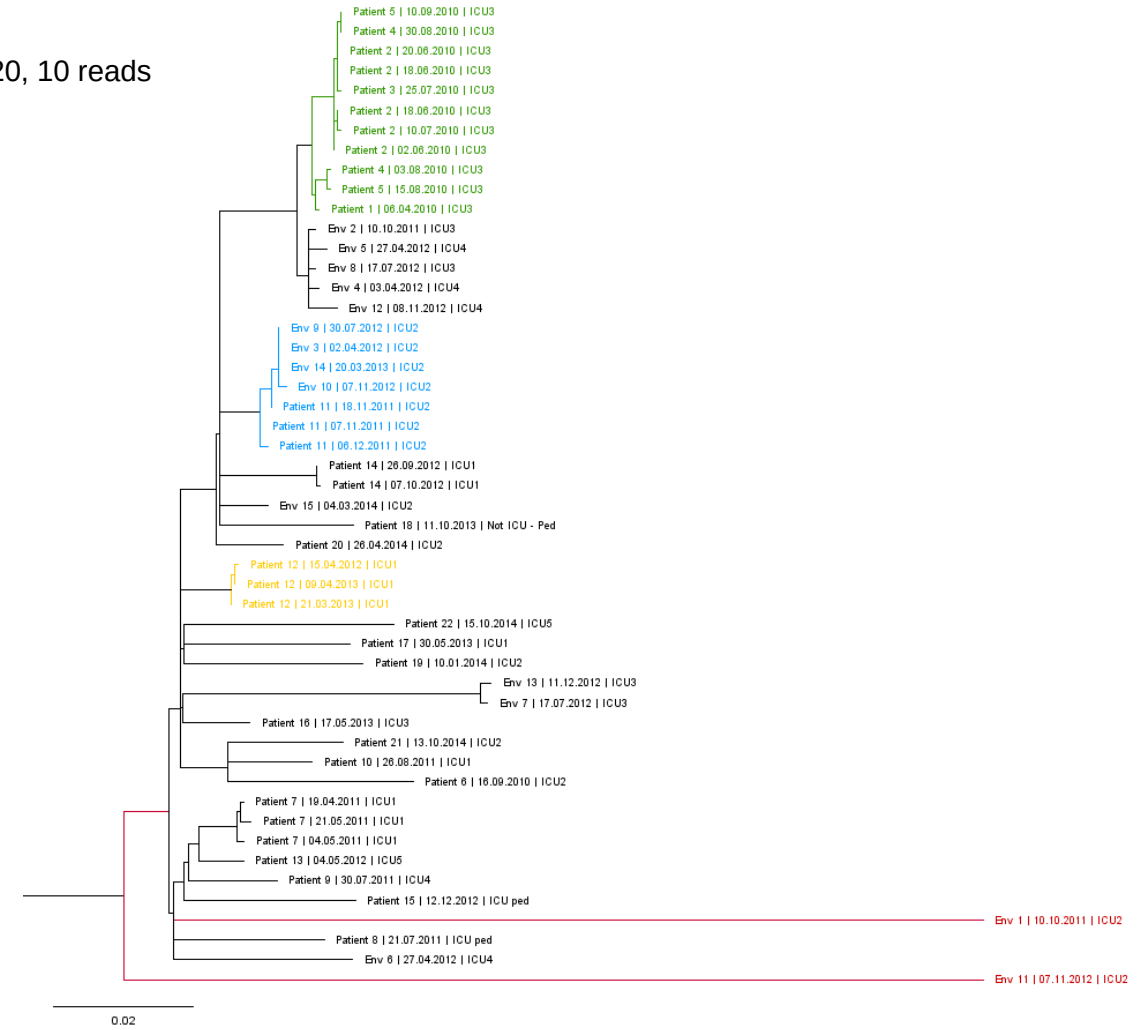
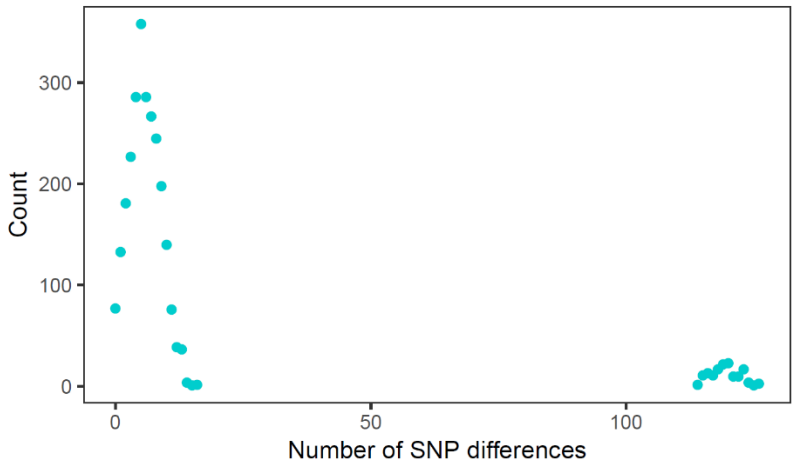


Figure 67. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacbBio reference with a mapping quality of 20 and with a minimum of 10 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red

(A)



(B)

(C)

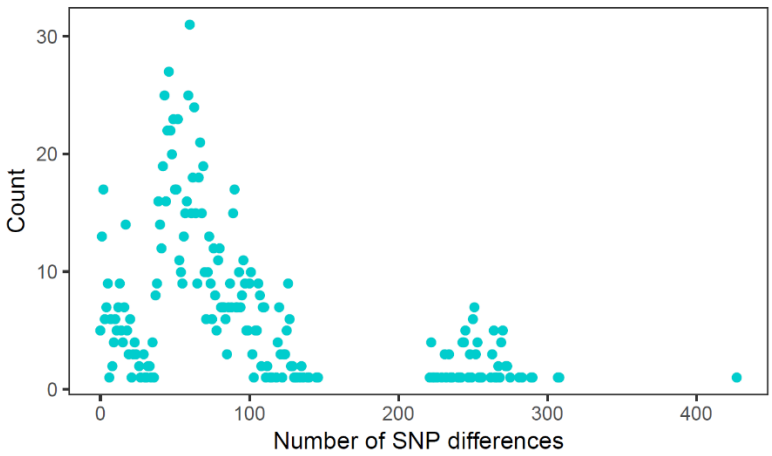


Figure 68. Frequency of number of SNP differences obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 10 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

DLST 1-18 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 20, 20 reads

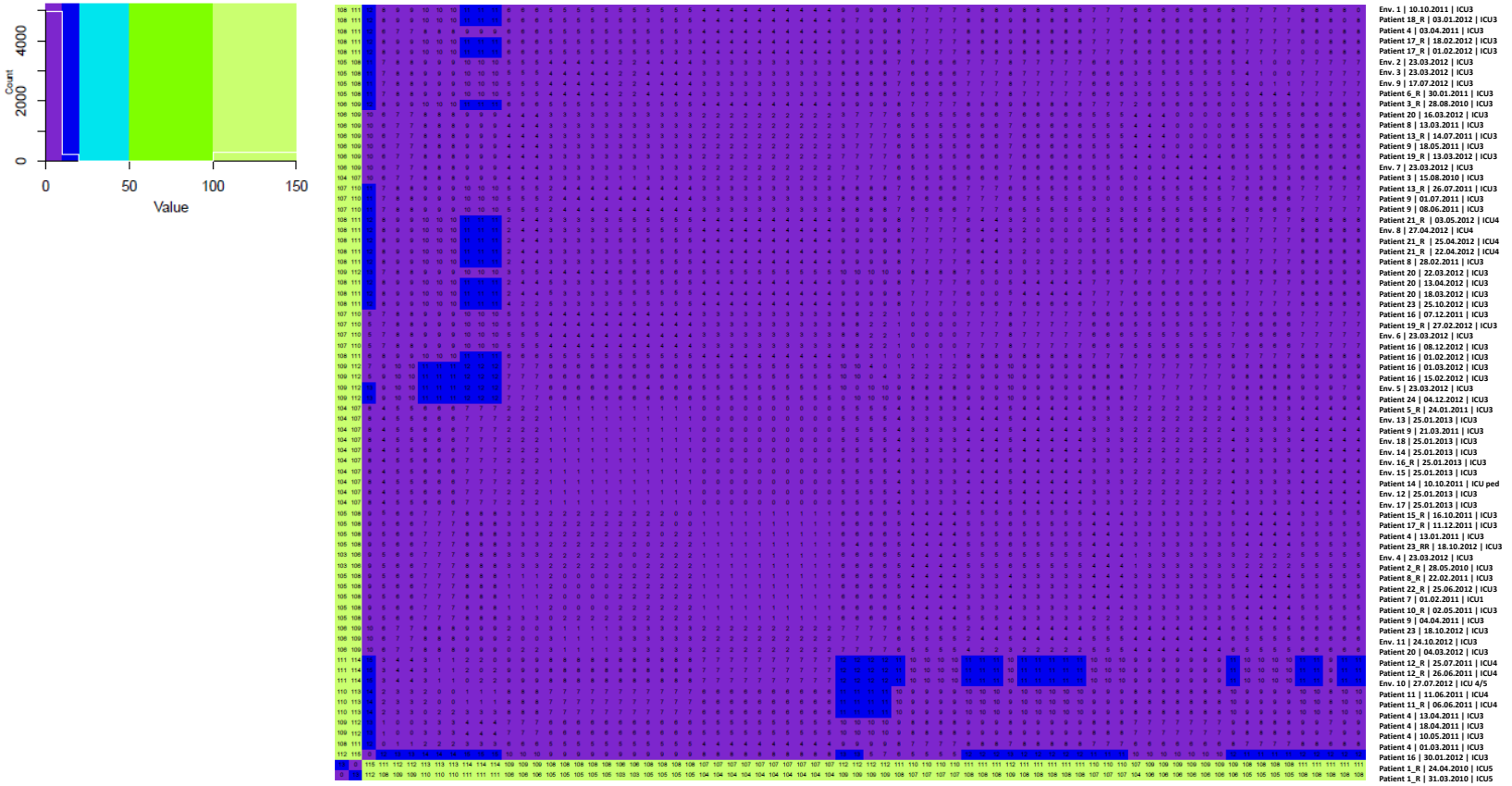


Figure 69. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Env. 1 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-18 phylogeny

Adapted methodology: PacBio reference, MQ 20, 20 reads

Figure 70. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology mapping against the PacBio reference with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

DLST 1-21 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 20, 20 reads

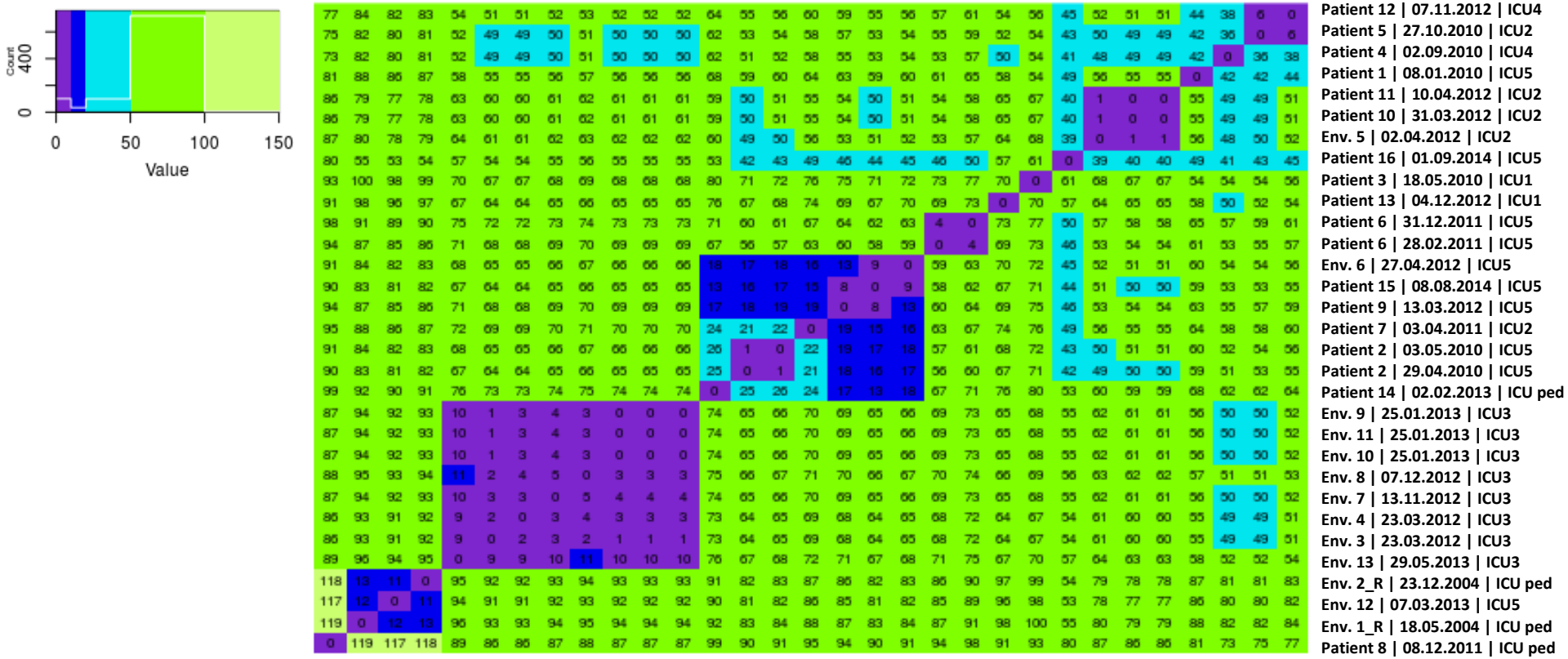


Figure 71. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 12 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PacBio reference, MQ 20, 20 reads

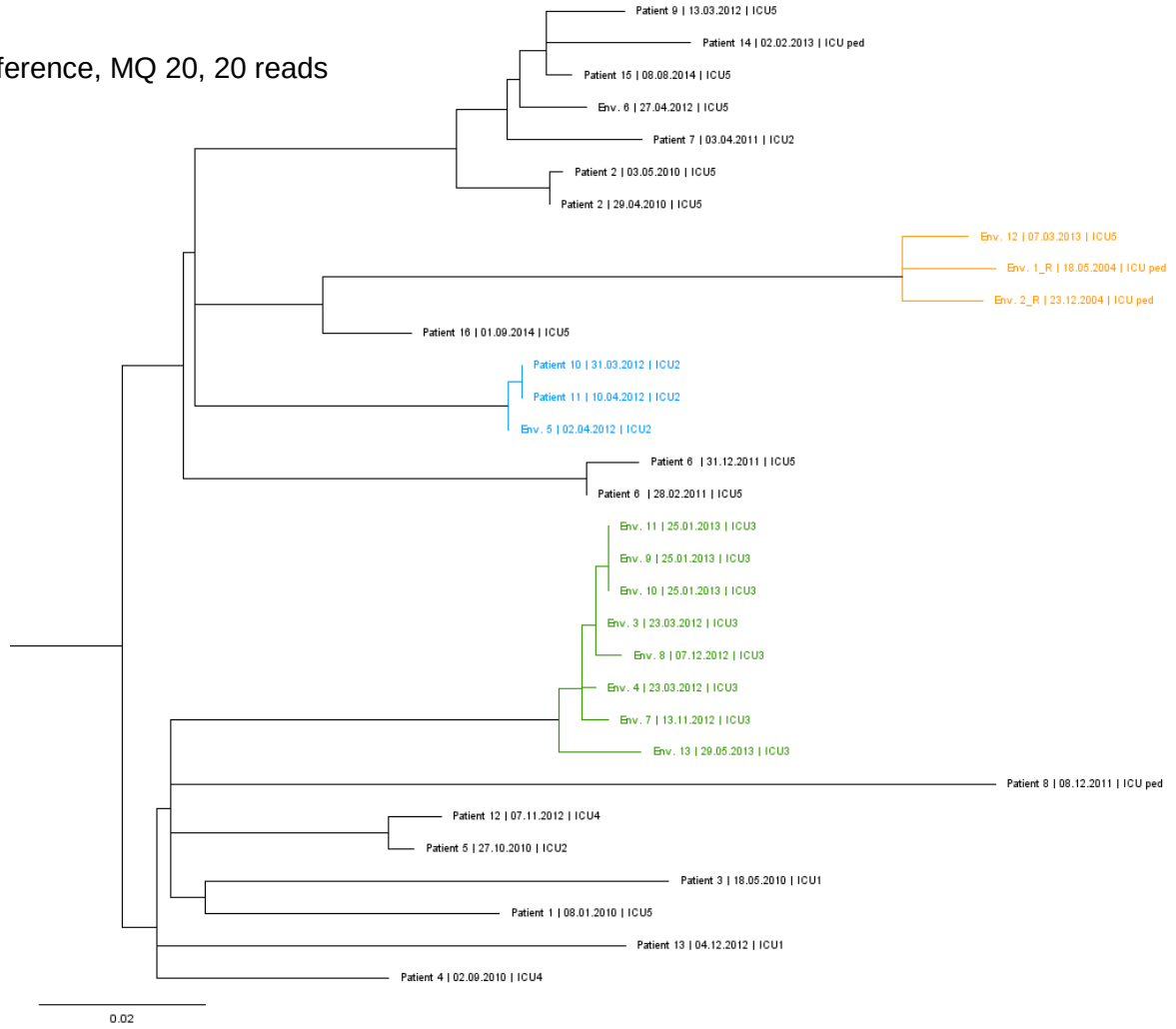


Figure 72. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacBio reference with a mapping quality of 20 and with a minimum of 20 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 20, 20 reads

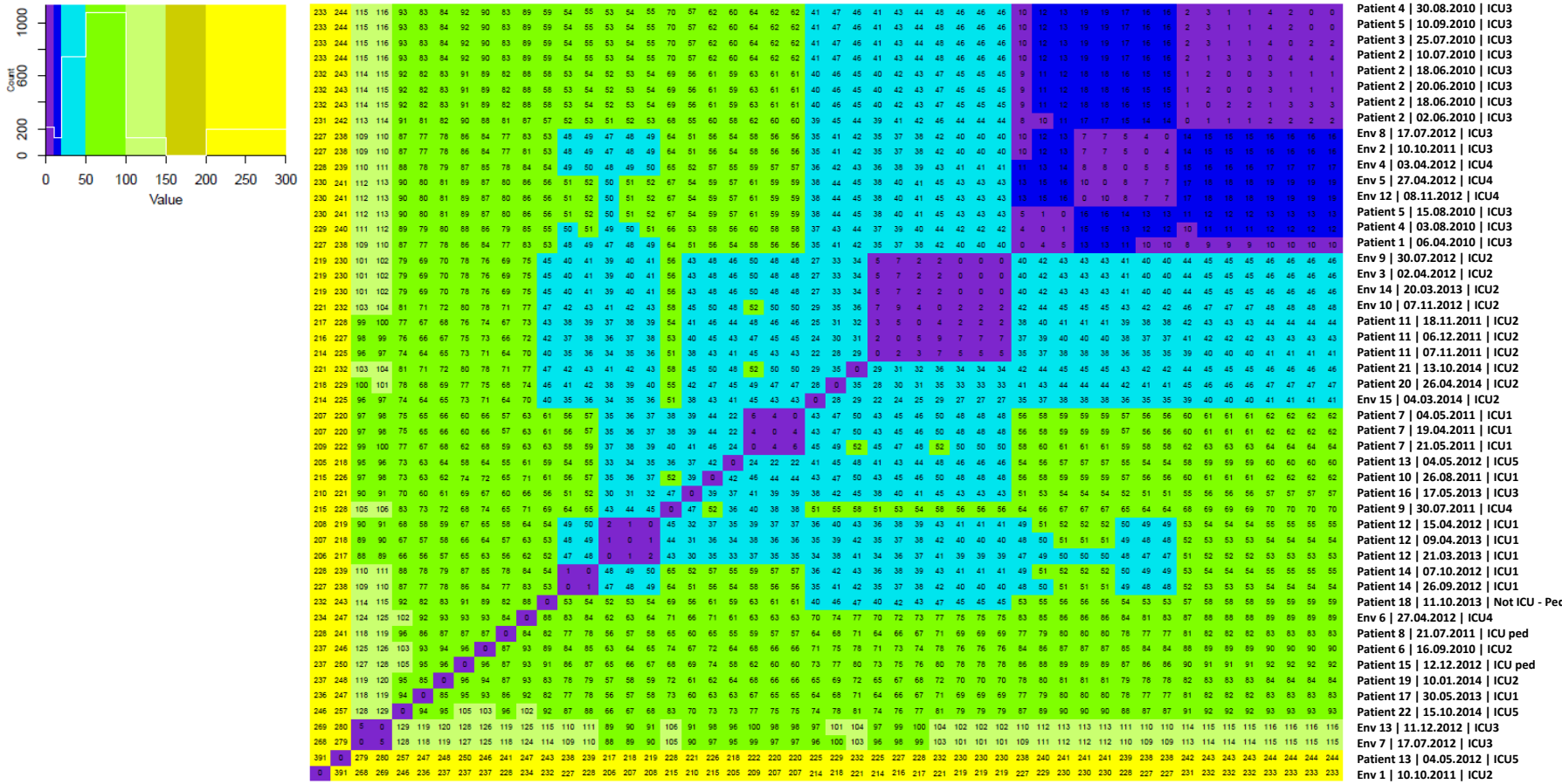


Figure 73. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Patient 4 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 300. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 6-7 phylogeny

Adapted methodology: PacBio reference, MQ 20, 20 reads

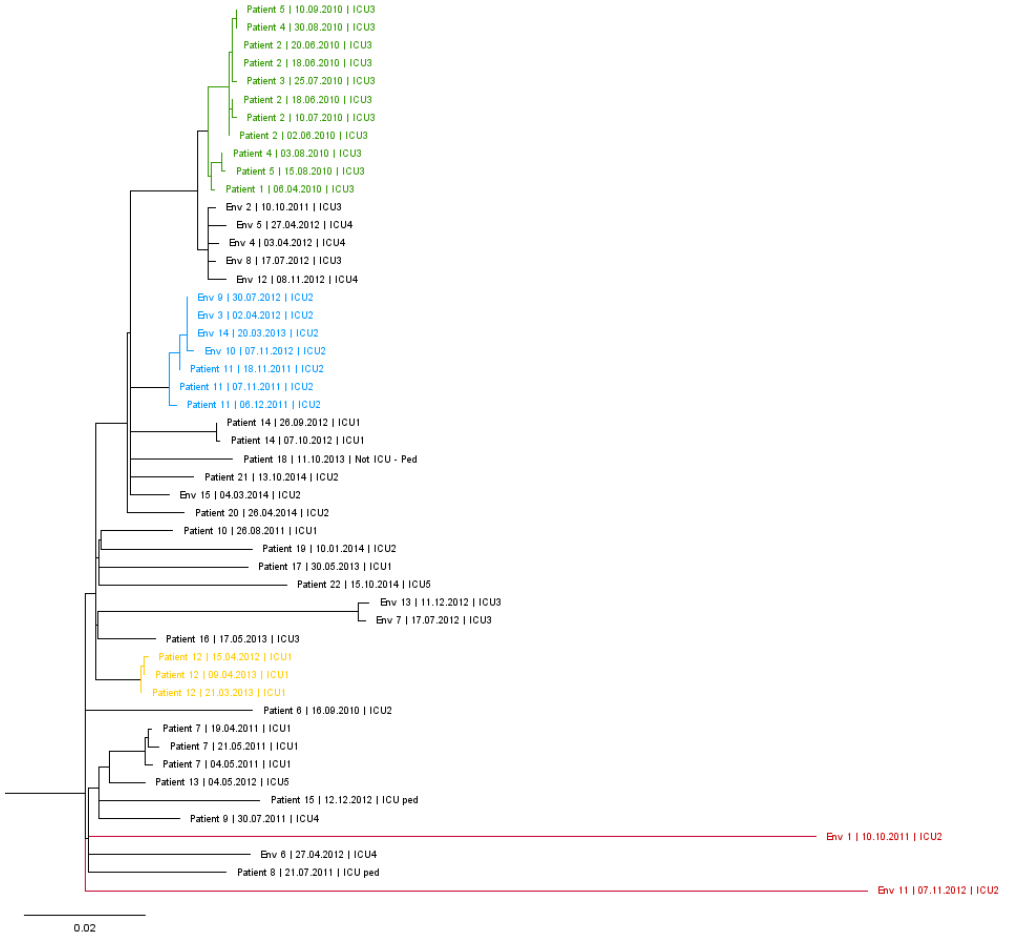


Figure 74. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacbBio reference with a mapping quality of 20 and with a minimum of 20 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red

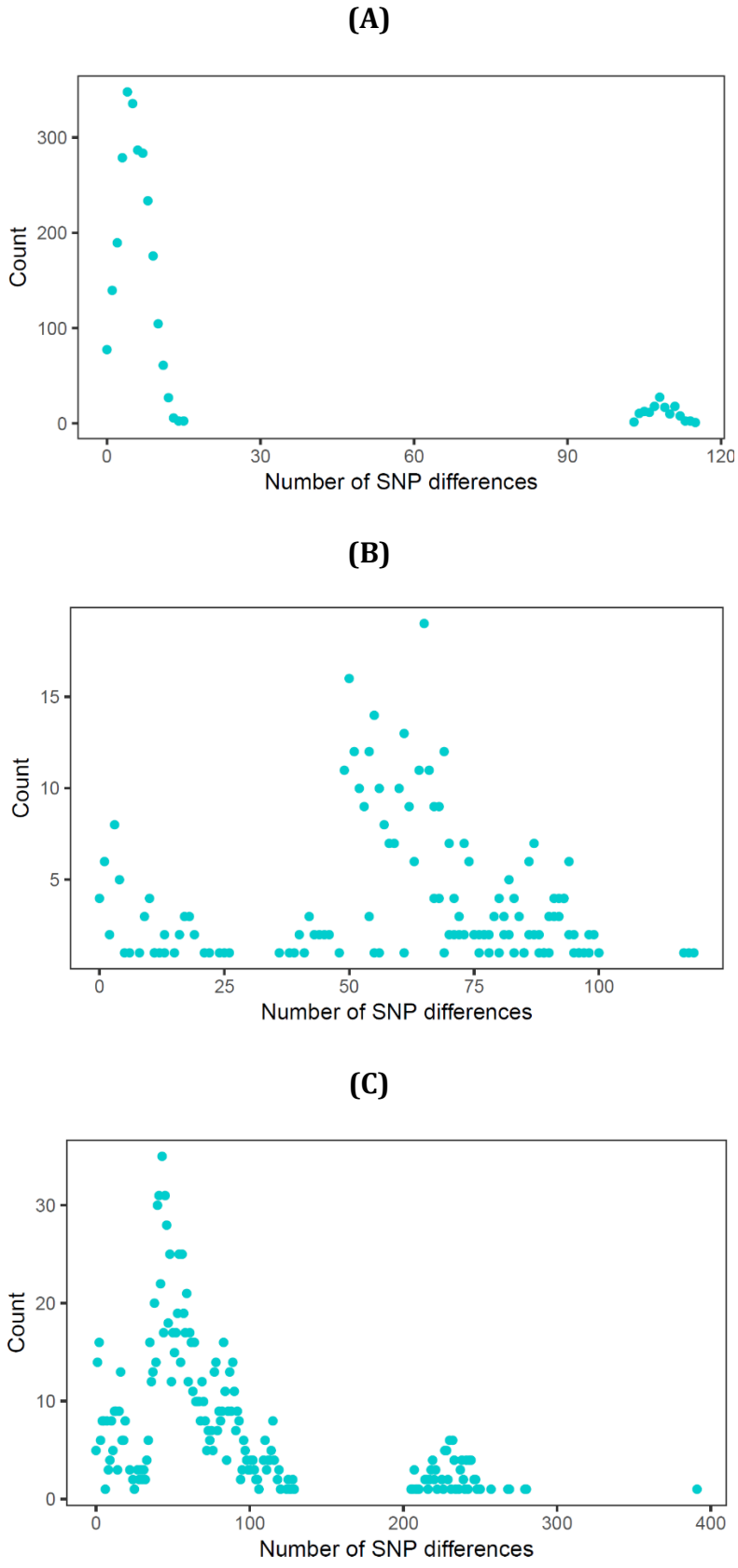


Figure 75. Frequency of number of SNP differences obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 20 and minimum of 20 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

DLST 1-18 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 60, 20 reads

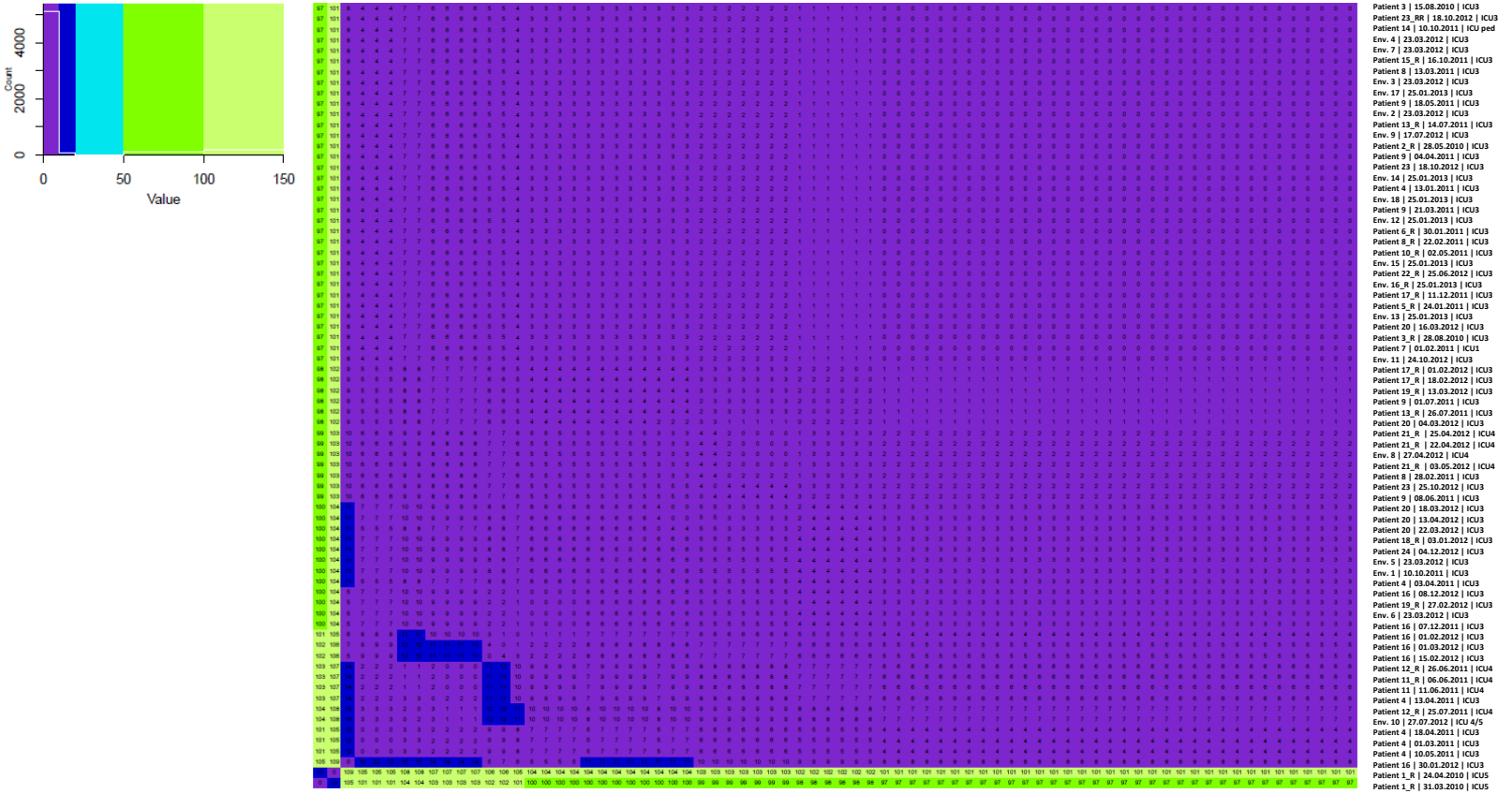


Figure 76. DLST 1-18 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Patient 3 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-18 phylogeny

Adapted methodology: PacBio reference, MQ 60, 20 reads

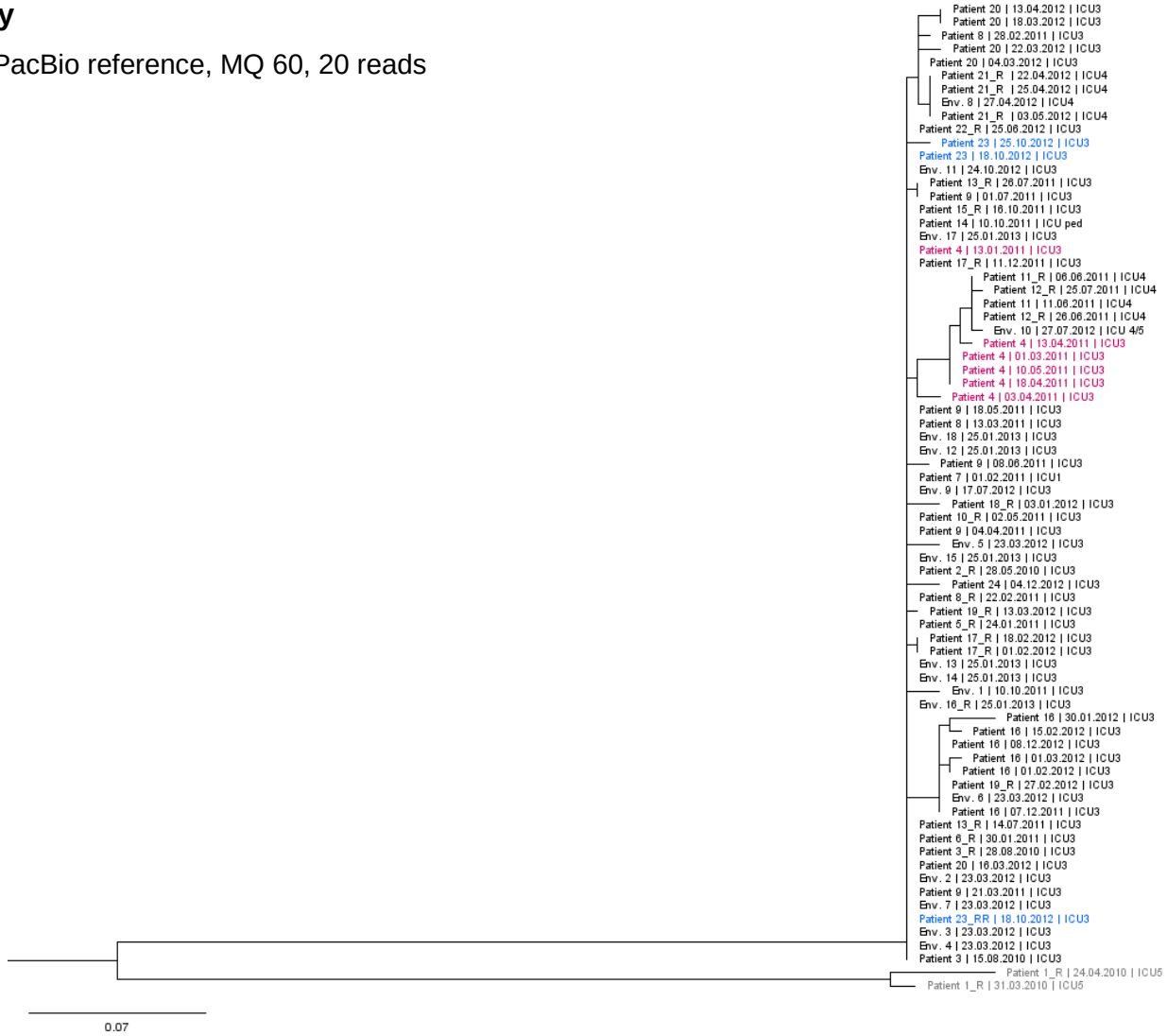


Figure 77. DLST 1-18 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology mapping against the PacBio reference with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Non-outbreak isolates belonging to Patient 1 are highlighted in grey. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

DLST 1-21 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 60, 20 reads

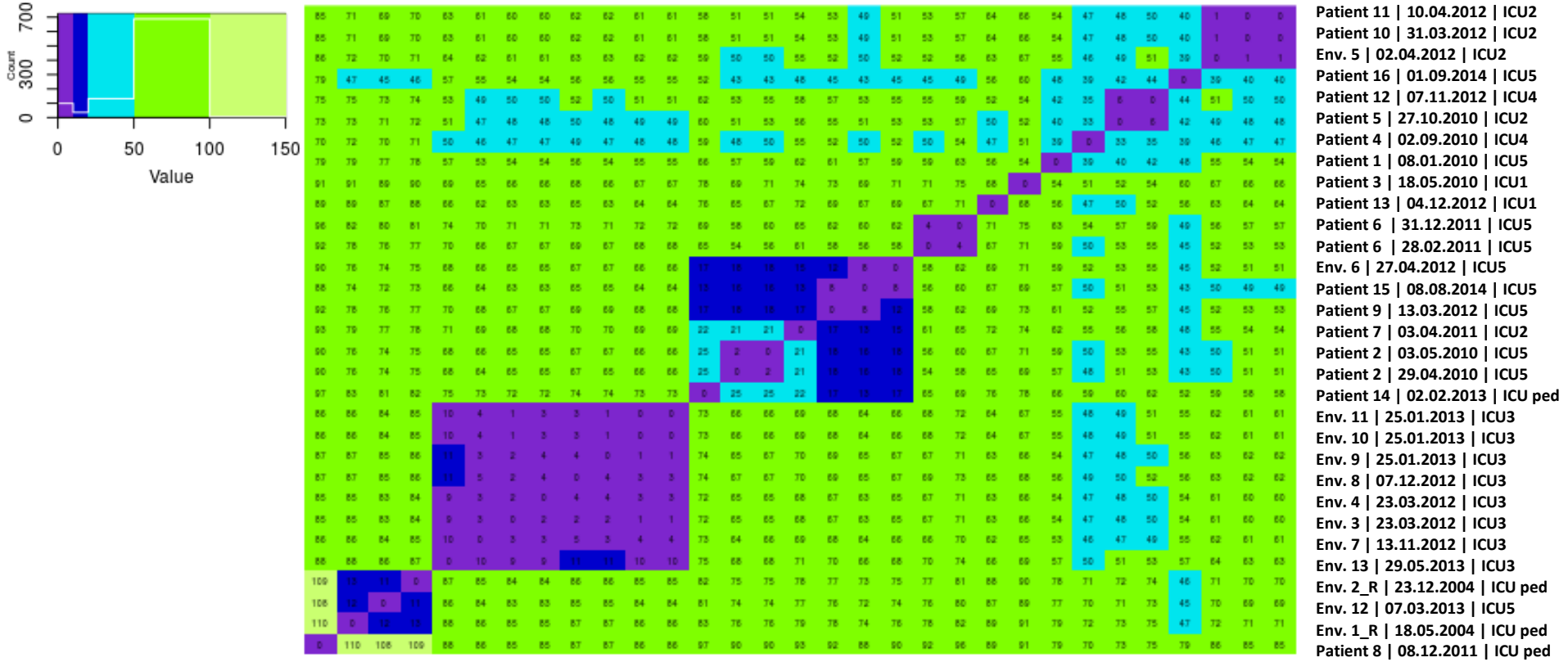


Figure 78. DLST 1-21 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 1-21 phylogeny

Adapted methodology: PacBio reference, MQ 60, 20 reads

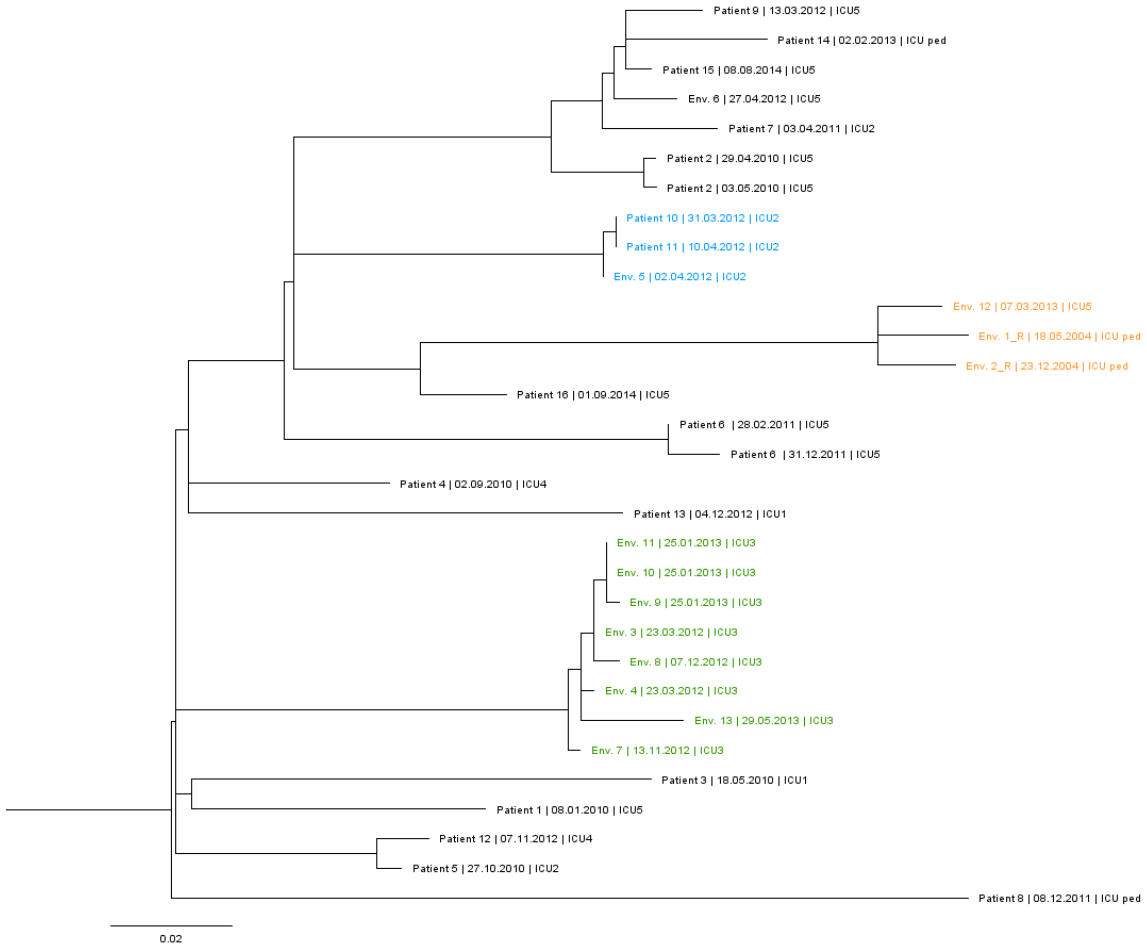


Figure 79. DLST 1-21 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacBio reference with a mapping quality of 60 and with a minimum of 20 reads to consider a SNP site. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

DLST 6-7 pairwise distance matrix

Adapted methodology: PacBio reference, MQ 60, 20 reads

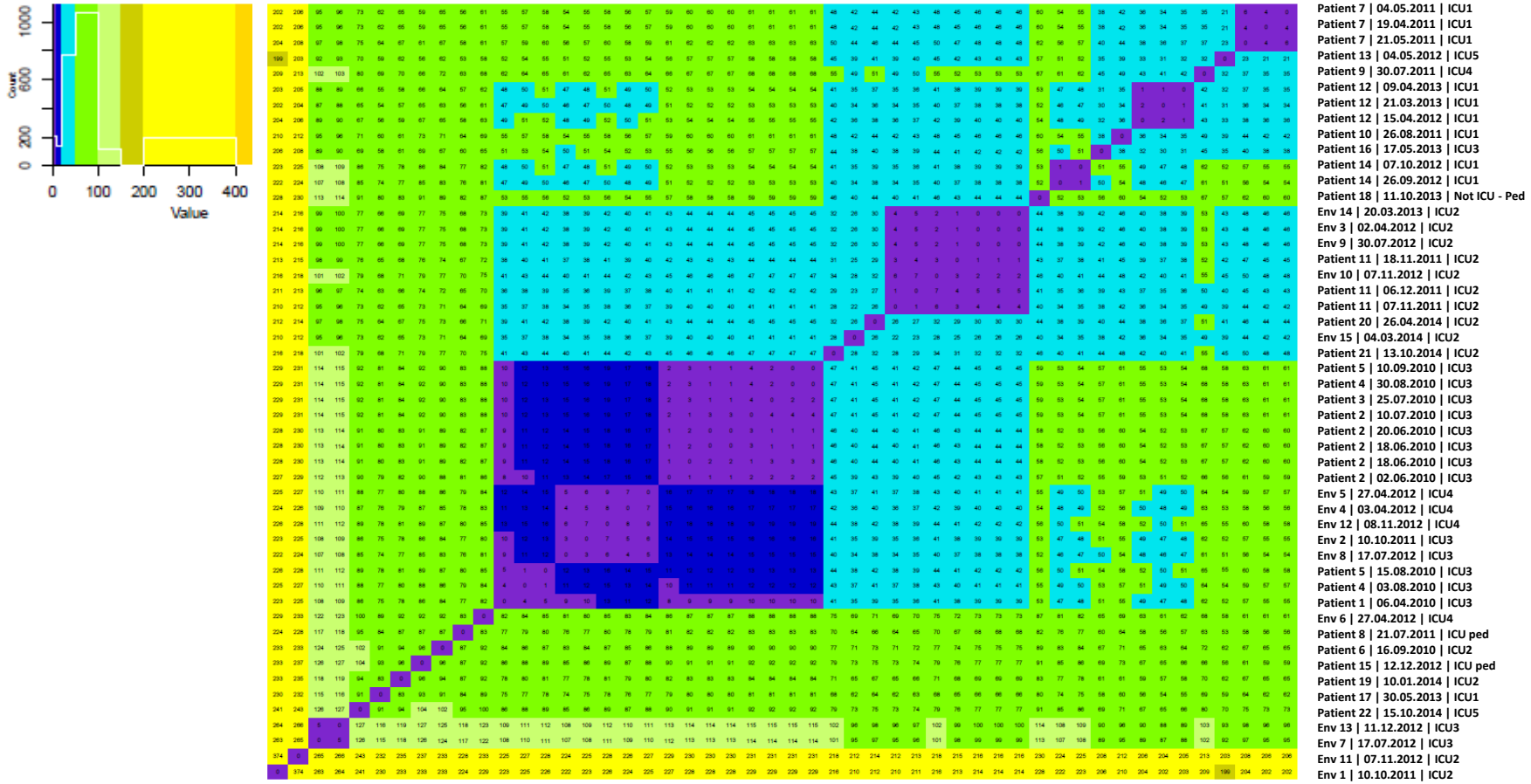


Figure 80. DLST 6-7 color heatmap showing pairwise genomic distances obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 60 and minimum of 20 reads to consider a SNP site. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Patient 7 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, 150, 200, and 400. A white line on the color legend plot pictures the frequency of each number of SNP differences.

DLST 6-7 phylogeny

Adapted methodology: PacBio reference, MQ 60, 20 reads

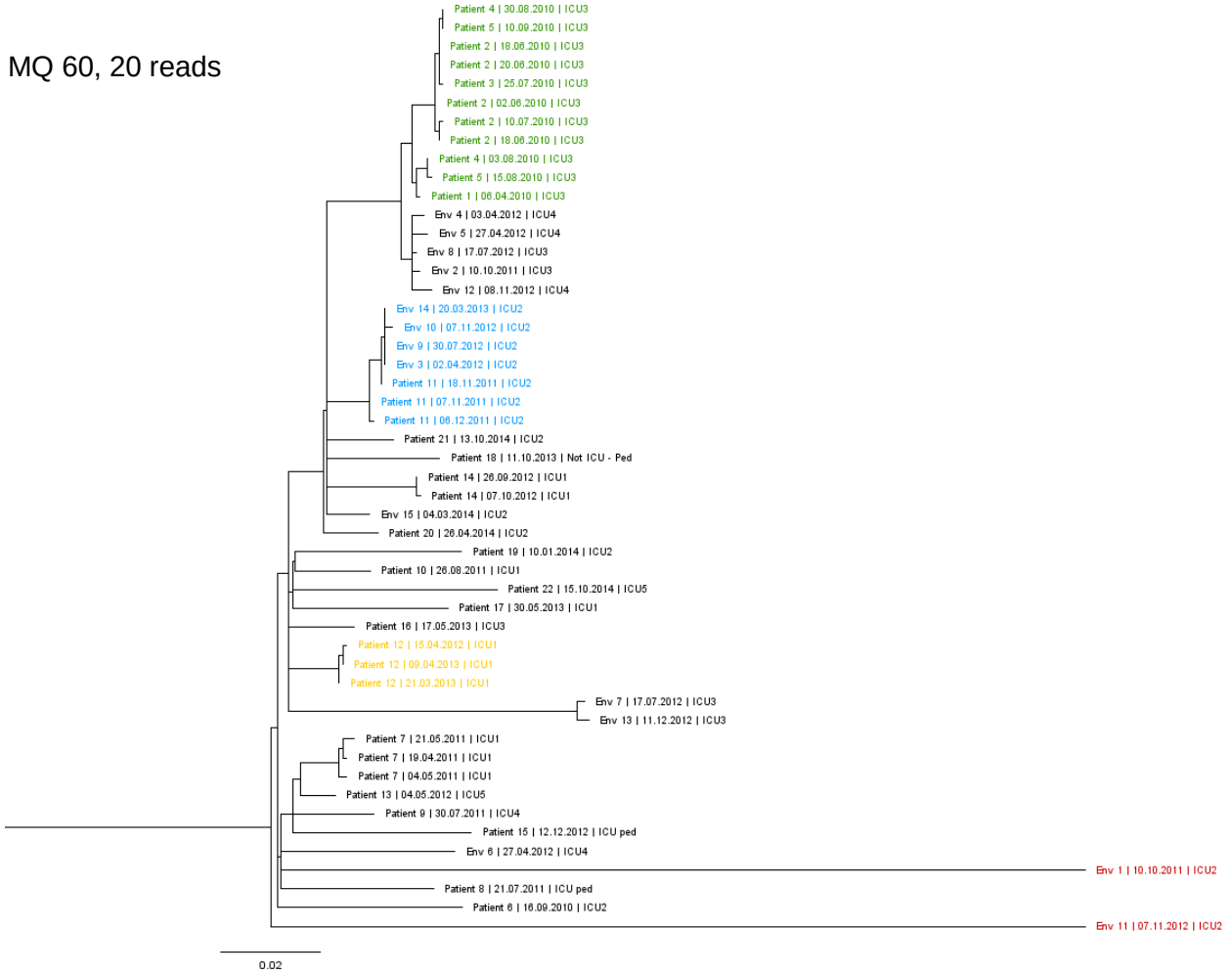
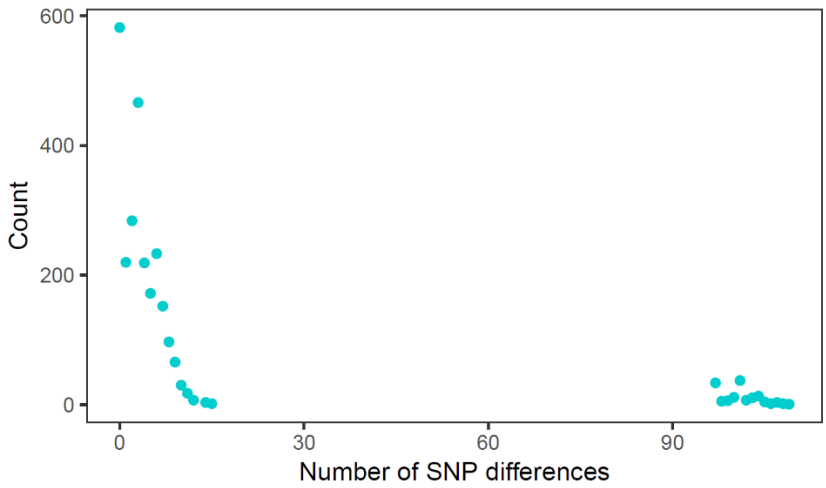
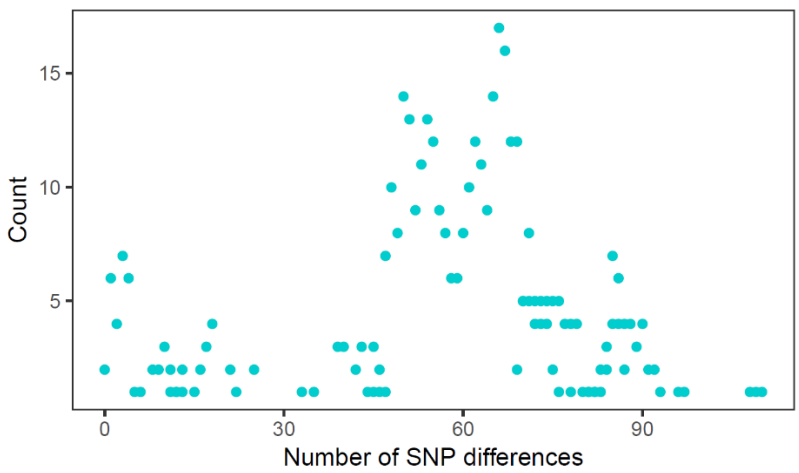


Figure 81. DLST 6-7 maximum likelihood tree based on the SNPs alignment obtained with the adapted methodology, mapping against the PacbBio reference with a mapping quality of 60 and with a minimum of 20 reads to consider a SNP site. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red

(A)



(B)



(C)

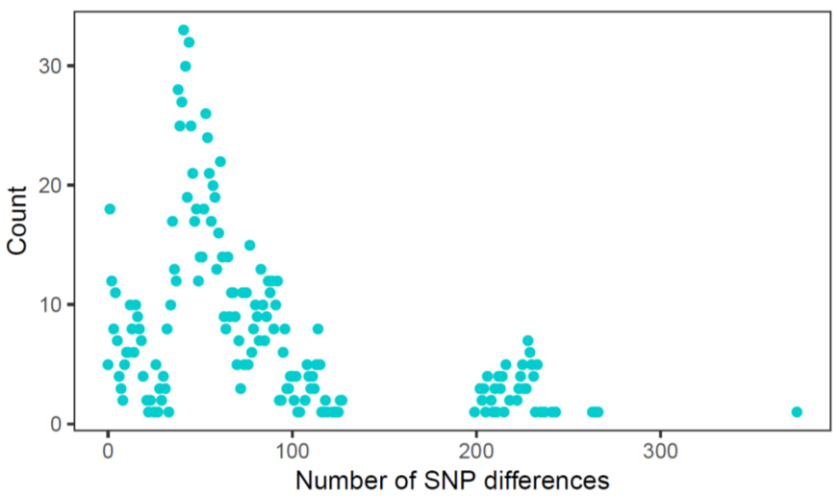


Figure 82. Frequency of number of SNP differences obtained with the adapted methodology, mapping against the PacBio reference with mapping quality of 60 and minimum of 20 reads to consider a SNP site, for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

Unveiling *Pseudomonas aeruginosa* epidemiology in intensive care units through combination of molecular typing and whole genome sequencing

(*Manuscript in preparation)*

Bárbara Magalhães[1], Benoit Valot[2], Mohamed M. H. Abdelbary[1], Guy Prod'hom[2], Gilbert Greub[3], Laurence Senn[1], Dominique S. Blanc[1]

[1]Service of hospital preventive medicine, Lausanne University Hospital, Lausanne, Switzerland,
[2]Chrono-environment, Franche-Comté University, Besançon, France,
[3]Institue of Microbiology, Lausanne University Hospital, Lausanne, Switzerland

Keywords: Whole genome sequencing, *P. aeruginosa*, molecular typing, epidemiology

1. Introduction

Pseudomonas aeruginosa is considered one of the main Gram-negative bacteria causing hospital acquired infections (1). In these settings, *P. aeruginosa* widely present in the environment and can be retrieved from different sources, like respiratory therapy equipment, antiseptics, soap, sinks, and hydrotherapy pools(2). This pathogen was also found to be part of the endogenous microbiota of 2.6 to 24% of the hospitalized patients (3, 4). Patients with compromised host defense mechanisms, such as neutropenia, severe burns, or cystic fibrosis, are particularly affected by this pathogen whose infections lead to high morbidity and mortality (5, 6). *P. aeruginosa* has been previously described as the second most common organism responsible for infections acquired in intensive care units (ICUs) (7)

P. aeruginosa population structure is consensually believed to be panmictic-epidemic (8-10), i.e. a superficially clonal structure with frequent recombination that creates new strains with unique genetic characteristics, in which occasionally highly successful epidemic clones arise. In addition, clinical isolates are indistinguishable from environmental isolates; and there are no specific clones related to a specific habitat selection (10).

Molecular epidemiological investigations have become indispensable for active surveillance of infection and detection of expanding disease outbreaks*. P. aeruginosa* possesses a very complex ecology and, for that reason, only powerful typing methods can give insight on the relatedness of strains, and consequently on the routes of colonization and/or infection (11). Pulsed-field gel electrophoresis (PFGE) has been considered the "gold standard" for DNA fingerprinting of *P. aeruginosa* (12-14). However, this method as several disadvantages, such as long analysis time, the use of expensive and specialized equipment, low intra- and inter-laboratory reproducibility and is labor-intensive, which make it not the optimal method to be used in a large investigation (14, 15). To overcome these limitations, alternative amplification-based molecular methods have been implemented as is the case of multi-locus sequence typing (MLST) which showed to be efficient in the study of the global population structure of *P. aeruginosa* (16). Another sequence-based method, double locus sequence typing (DLST), based on partial sequencing of two highly variable loci has been successfully used to investigate the epidemiology of *Staphylococcus aureus* and *Pseudomonas aeruginosa* (1, 17-19).More recent studies on the *P. aeruginosa* evolution and dissemination in hospital settings have been conducted by recurring to whole genome sequencing (WGS) (20-22). This method enables the analysis of the complete genome of bacterial isolates, distinguishing strains at the single nucleotide level.

An increase in *P. aeruginosa* incidence was observed in the ICUs of the Uiversity Hospital of Lausanne. Clinical and environmental isolates retrieved from 2010 to 2014 were typed using DLST (1). Three major DLST types, i.e. comprising the highest number of patients, where identified: DLST 1-18, DLST 1-21, and DLST 6-7. DLST 1-18 was previously reported as the cause of an outbreak in the burn unit (23). However, DLST 1-21 and DLST 6-7 showed sporadic occurrence with only few cases of possible transmission between patients. The discriminatory power of whole genome sequencing (WGS) was used to further investigate these three major DLST clusters.

2. Material and Methods

Bacterial isolates and molecular typing. *P. aeruginosa* isolates were collected from patients hospitalized in the five ICUs of the University Hospital of Lausanne over a five-year period. From 2010 to 2014, clinical and environmental isolates were typed by the double locus sequence typing (DLST) method (1) previously developed by our group. Three major DLST clusters, i.e. clusters with the highest number of patients, were further analysed in this study: DLST cluster 1-18 (24 patients), 6-7 (22 patients), and 1-21 (16 patients). At least one isolate was selected per patient. If several isolates were collected from one patient, only isolates sampled 15 days apart were selected, unless they belonged from different sample types. All environmental isolates from the three DLST clusters (mainly from sink traps) were considered. A total of 74 DLST 1-18 isolates (55 clinical and 19 environmental), 50 DLST 6-7 isolates (38 clinical and 12 environmental), and 31 DLST 1-21 isolates (18 clinical and 13 environmental) were selected for whole genome sequencing.

Epidemiological investigation. Epidemiological data (unit and room of hospitalization, dates of ICU admission and discharge, and clinical diagnosis) was retrieved from the hospital databases and used to construct epidemiological maps and annotate the phylogenetic trees. Epidemiological links between patients or environment were identified as: (i) patients hospitalized during overlapping periods in the same ICU, or (ii) patients showing an identical DLST type with an environmental sample isolated in the same unit during the period of the study.

DNA extraction and whole genome sequencing. We extracted genomic DNA from a 5ml Lysogenic Broth (LB) culture, acquired from single colonies and incubated to reach an early exponential phase, using the GenElute bacterial genomic DNA kit (SIGMA-ALDRICH,

St. Louis, MO, USA). Whole genome sequencing was performed on 155 *P. aeruginosa* clinical and environmental isolates by the Lausanne Genomic Technologies Facility (GTF, University of Lausanne). The sequencing libraries were prepared using the Nextera DNA Library Preparation Kit (Illumina, Inc., San Diego, CA, USA) for 100-bp paired-end sequencing runs on Illumina HiSeq 2500, aiming for a 100-fold coverage.

SNPs and phylogenetic analysis. Isolates' sequence type (ST) was assigned from the short reads data by the Short Read Sequence Typing 2 (SRST2) software. A first step of subsampling the number of raw reads to reach the lower read depth observed (70x) was added to provide comparable accuracy in the posterior analysis, as well as to reduce mapping time. Complete reference genomes were created by sequencing the first collected clinical isolate of each ST with both PacBio and Illumnina HiSeq technologies. The subsampled reads were then mapped against their respective complete reference genome with BWA-MEM. Variant calling was performed with FreeBayes with a minimum mapping quality of 60. Putative phages found with PHASTER along with repeat regions and potential recombination regions detected with an in-house script, were excluded from the genome alignment.

A maximum likelihood tree was constructed from the final core SNPs alignment using the PhyML algorithm implemented in Seaview version 4.6.1 (24). Tree visualization was done with FigTree version 1.4.3. Detailed methodology is present in Supplementary data 1.

3. Results

Different epidemiology of the three DLST clusters. To infer possible epidemiological links between patients of the same DLST cluster and/or between patients and environmental isolates, the hospitalization period, the ICU where the hospitalization occurred, and the ICUs environmental sampling of *P. aeruginosa* were investigated and

are schematically represented in Figures 2.23. DLST cluster 1-18 was previously considered responsible for an outbreak in the burn unit from 2010 to 2012 (23). From the 24 patients harbouring this DLST type, 18 were hospitalized in the burn unit (ICU 3), and six in other ICUs. Several epidemiological links were found between the patients hospitalized in the burn unit which shared the same hydrotherapy shower room, during overlapping hospitalization periods. The first patient observed harbouring this DLST type was hospitalized in ICUs 1 and 5 and was not epidemiologically linked to other patients infected with the same type. Links between environmental isolates, mainly from the shower room and sink traps, and patients hospitalized in the same ICU were also found.

Only two epidemiological links were identified for DLST cluster 1-21; one between two patients hospitalized in the same ICU (ICU 2), and one between those patients and an environmental sample retrieved from a sink trap in the same ICU. The remaining patients were dispersed through the six ICUs during the study period, except in 2013 when no patient was found to be colonized or infected with this DLST type (Figure 1).

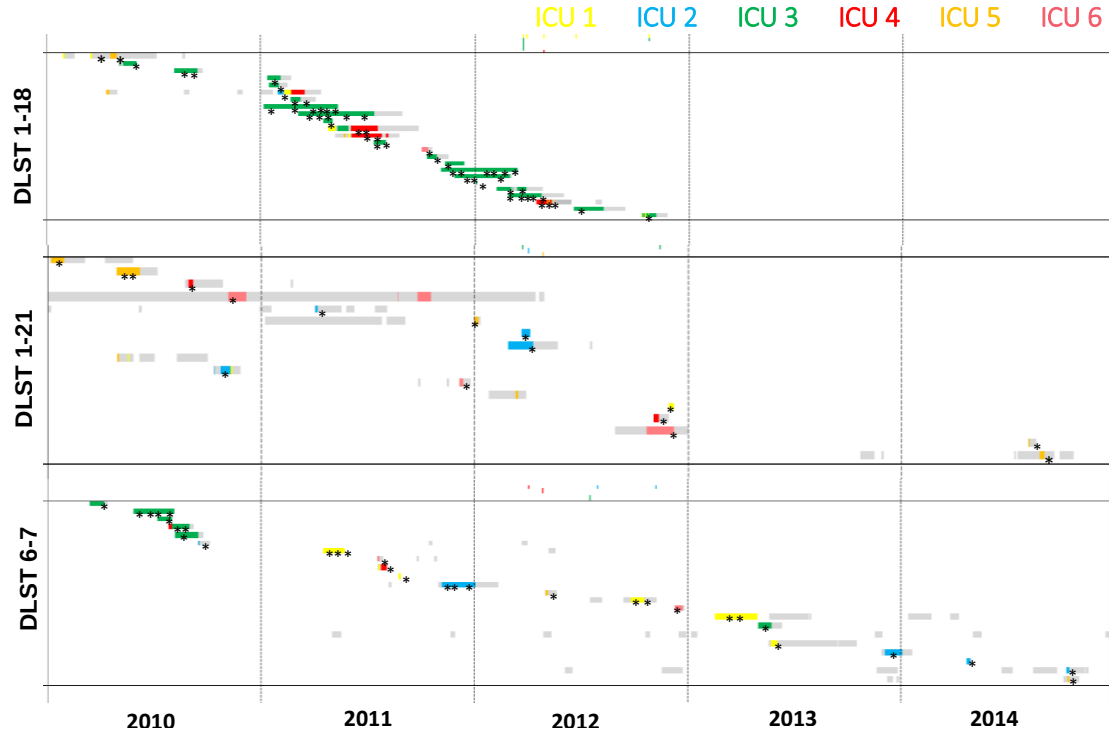


Figure 1. Epidemiological maps of the three different DLST types. The first panel corresponds to patients harbouring DLST 1-18 (N=24), the second to DLST 1-21 (N=16), and the third to DLST 6-7 (N=22). Each line represents the hospitalization period of each patient from 2010 to 2014. Units where patients were hospitalized are differentiated by colors. Stars represent the first isolation of *P. aeruginosa* for each patient.

Such behaviour suggests DLST cluster 1-121 was not considered to be the cause of a *P. aeruginosa* outbreak.

Three epidemiological links were found between DLST cluster 6-7 patients hospitalized in the burn unit, in 2010. From 2011 to 2014, no epidemiological links were suspected as patients were not hospitalized in the same ICU during overlapping periods of time. In addition, no epidemiological links between patients and environmental sources were observed. Similarly to DLST cluster 1-21, this DLST type occurred sporadically throughout the study period and was not responsible for an outbreak.

Different DLST clusters belonged to different sequence types. Among the different genotyes, DLST cluster 1-18, 1-21, and 6-7 comprised the highest number of patients and were chosen for posterior analysis with WGS. Although DLST allows inter laboratory comparison of genotypes, the universal standard of multilocus sequence typing (MLST) is still widely used for strain comparison. Therefore, the Illumina HiSeq raw reads were used to identify the STs present in the isolate collection. MLST results defined DLST 1-18, 1-21, and 6-7 isolates as STs 1076, 253, and 17, respectively, confirming the previously documented similar discriminatory power of both methods (25).

WGS confirmed the outbreak caused by DLST cluster 1-18. Patient 1, which was not hospitalized in the burn unit and had no epidemiological link with the outbreak, clustered far apart from the remaining isolates with a maximum of 120 SNPs between a pair of isolates (Figure 2 and 3). Several subclades were identified comprising both clinical and environmental isolates, however most of the outbreak isolates were closely related with less than 10 SNP differences. Patient 4 isolates from ICU3 and Patient 11,12, and environmental isolate 10 from ICU4 were integrated in the same subclade. These isolates acquired from ICU4 and an isolate from Patient 4 (date of collection: 13.04.2011) showed a slightly higher number of differences in relation to the remaining isolates (<16 SNPs). Isolates retrieved from the same patient, although some were

separated on the phylogenetic tree, were genetically different by only less then 10 SNPs differences, e.g. Patient 4 (<7 SNPs) and Patient 23 (0-2 SNPs). The number of SNP differences count (Figure 4) demonstrated that the maximum of differences observed between a pair of isolates is 16 and the minimum is zero, except for isolates belonging to Patient 1 (120 SNPs).

DLST cluster 1-21 and 6-7 occurred sporadically in the ICUs. Two clades and several subclades with high SNP differences between them reinforce the premise that DLST 2-21 was not responsible for an outbreak. Two isolates from two patients (zero SNPs) and one isolate from the environment (Figure 5 and Figure 6, in blue,) showed one SNP difference between them, confirming the previously suspected epidemiological link. Only six SNP differences were found between two isolates of two patients hospitalized in different ICUs (ICU2 and ICU4), retrieved two years apart. A low number of SNPs (<14 SNPs) was observed between environmental isolates retrieved ten years apart (Figure 6, in orange). Environmental isolates retrieved from different sink traps in the burn unit (ICU3) clustered together with less than 11 SNP differences. Isolates belonging to the same patient were closely related with a maximum of 4 SNPs between them. SNPs count for this DLST cluster demonstrated a high variability of the number of SNPs found, ranging from zero to 116 SNPs between a pair of isolates (Figure 4).

DLST 6-7 clades and subclades showed a high number of SNP differences (Figure 7). One subclade (Figure 8, in green) was composed by isolates retrieved in the burn unit with less than 13 SNPs differences. These isolates belonged to patients for which epidemiological links were suspected. Four environmental isolates retrieved from both the burn unit and ICU4 were closely related to the burn unit cluster (10-19 SNPS). A second subclade was constituted by closely related isolates from Patient 11 sampled in ICU2 and environmental isolates from the same ICU (0-7 SNPs). All isolates recovered from the same patient clustered together with only a few SNPs differences (1-6), e.g.

Patient 12 (1-2 SNPs) (Figure 8, in yellow). Long branches with more than 200 SNPs associated with two isolates from the environment (Env 1, 11) were detected. Most isolates had 50 to 150 SNP differences between them, with a maximum of 429 SNPs and a minimum of 2 SNPs (Figure 4).

4. Discussion

In this study we give insight to the epidemiology of *P. aeruginosa* in the ICUs of the University Hospital of Lausanne by combining a molecular typing method with WGS. The three investigated DLST types sowed different epidemiological behaviours during this study period. Most of DLST cluster 1-18 patients were hospitalized in the burn unit during overlapping periods of time. As *P. aeruginosa* is capable to survive on wet surfaces such as sinks, sink traps, pipes, and hydrotherapy equipment; several nosocomial outbreaks have been associated with these specific reservoirs (26). DLST 1-18 environmental isolates retrieved from shower mattresses and sink traps from the hydrotherapy room support the assumption of an environmental source of infection. The high number of epidemiological links between patients, along with the wide presence of this DLST type in the environment of the burn unit, helped to previously determine this cluster as responsible for an outbreak with an environmental source (23). However, the first patient detected with DLST 1-18 (Patient 1) was not considered an outbreak patient as since it was not epidemiologically linked to the other patients.

In 2010, DLST type 6-7 was responsible for a small outbreak in the burn unit comprising five patients. From 2011, both DLST 1-21 and 6-7 occurred sporadically throughout the rest of the study period with only one suspected epidemiological link found for DLST 1-21 isolates (between patients and environment). This behaviour may be explained by a major role of this types' prevalence in our ICUs environment which lead to sporadic patient infection. Nonetheless, one limitation of this study relies on the

insufficient environmental sampling information until 2012. A more frequent and regular sampling throughout the four years study would have helped to discover probable epidemiological links between infected patients and environmental sources.

MLST results acquired from the isolates' raw reads divided the dataset in three different STs; ST 1076, ST253, and ST17. This division was exactly concordant with the attribution of types performed with DLST; DLST 1-8, DLST 1-21, and DLST 6-7, respectively. Such findings confirmed the previously documented similar discriminatory power of both methods (25). ST253 belongs to the clinical and international well described clonal complex (CC) PA14, and ST17 was previously reported as part of the clonal complex C, both CCs being the worldwide most abundant clonal complexes in the *P. aeruginosa* population (27).

By combining DLST and epidemiological data it was possible to determine three genotypes with different behaviours in our ICUs. However, DLST was not discriminatory enough to confirm possible cases of transmission between patients and between patients and the environment, or to define a probable source of infection. WGS helped to group the DLST 1-18 outbreak isolates with less than 10 SNP differences between them, while excluding Patient 1 as part of the outbreak, which was inferred by epidemiological data but not by DLST typing. Environmental isolates retrieved from sink traps and shower mattresses on the hydrotherapy room clustered with the outbreak isolates (<10 SNPs) which can indicate them as possible sources of infection.

Analysis of DLST 1-21 WGS data confirmed the suspected epidemiological link between isolates retrieved from ICU2. In addition, it considered as closely related, isolates for which no epidemiological links were suspected. For instance, isolates sampled from the burn unit were related with less than 11 SNPs; environmental isolates sampled 10 years apart were related with 11 to 14 SNPs; two isolates from two patients collected 12

years apart had six SNP differences. These values are lower than expected when considering the long time between isolate sampling, and considering that isolates retrieved from the same patient, weeks apart, had close number of SNP differences (0-4 SNPs). One explanation can be the slower evolution of *P. aeruginosa* isolates in the environment of ICUs which then lead to patients being infected with genetically identical strains.

Lastly, a small DLST 6-7 outbreak between patients hospitalized in the burn unit in 2010 was confirmed by WGS (0-13 SNP differences). A subclade of ICU2 clinical and environmental isolates with zero to seven SNP differences suggests a possible transmission between patient and the environment that was not questioned with the epidemiological data. Interestingly, two environmental isolates were associated with long branches. One reason for the occurrence of these long branches is the long branch attraction phenomenon: phylogenetic artefact when distantly related lineages are erroneously considered closely related solely because they have both undergone a large amount of molecular change (28). Another reason could be that these are hypermutator isolates as a response to environmental selection (29). A way to assess the latter would be investigated the presence of genes coding for the methyl-directed mismatch repair (MMR) system proteins in this DLST type genome.

Although WGS costs are decreasing, its implementation as a routine surveillance method for *P. aeruginosa* still comes at a higher price per isolate than the currently used DLST. Additionally, analyses of WGS data requires a certain level of bioinformatic expertise that is not always available in all epidemiology laboratories (30). Thus, recurring to DLST as a first-line molecular typing tool for screening of cases important to be analysed with the discriminatory power of WGS would culminate in a accurate and cost-efficient typing strategy.

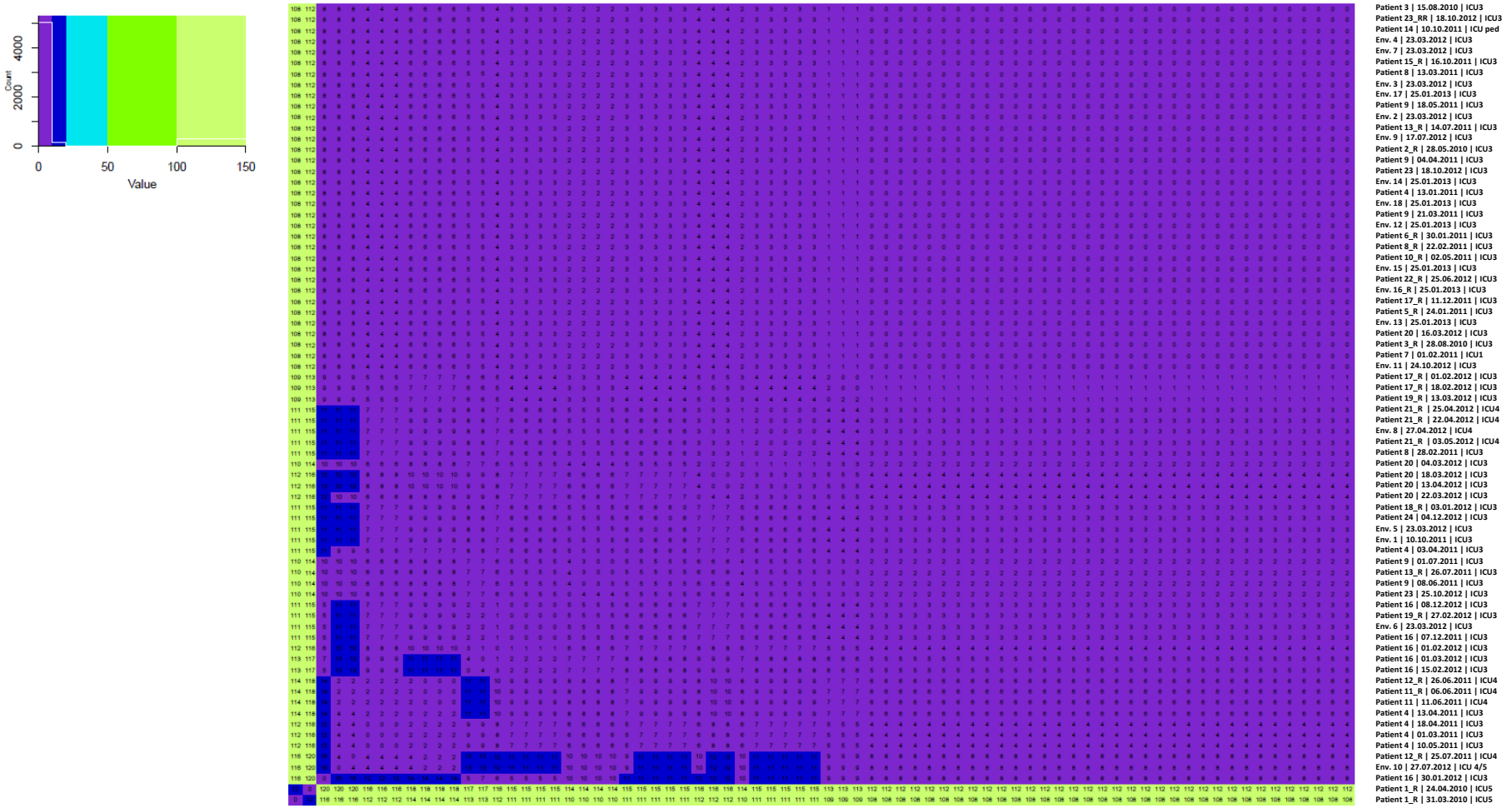


Figure 2. DLST 1-18 color heatmap showing pairwise genomic. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 1 (first isolate) to Patient 3 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

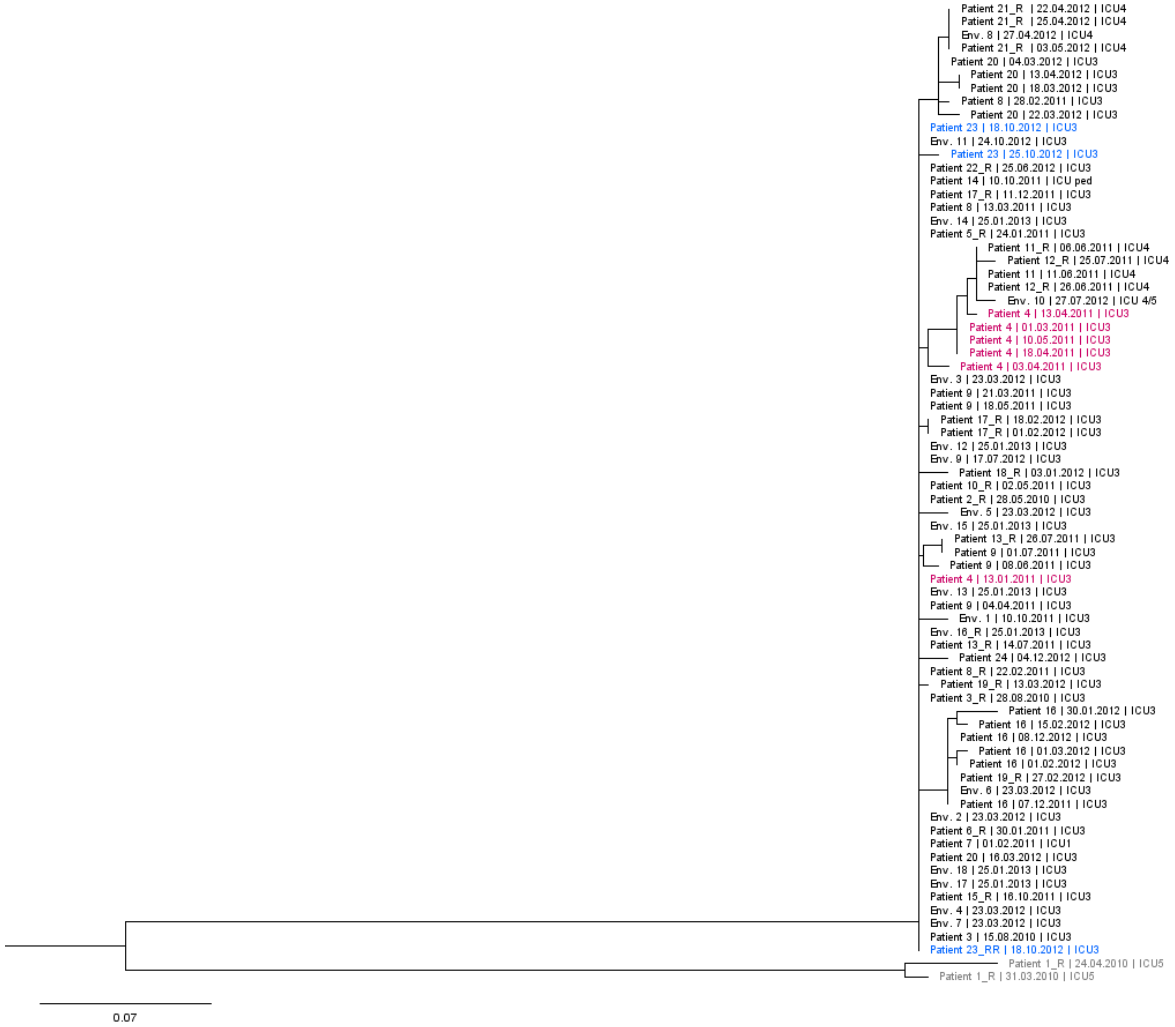
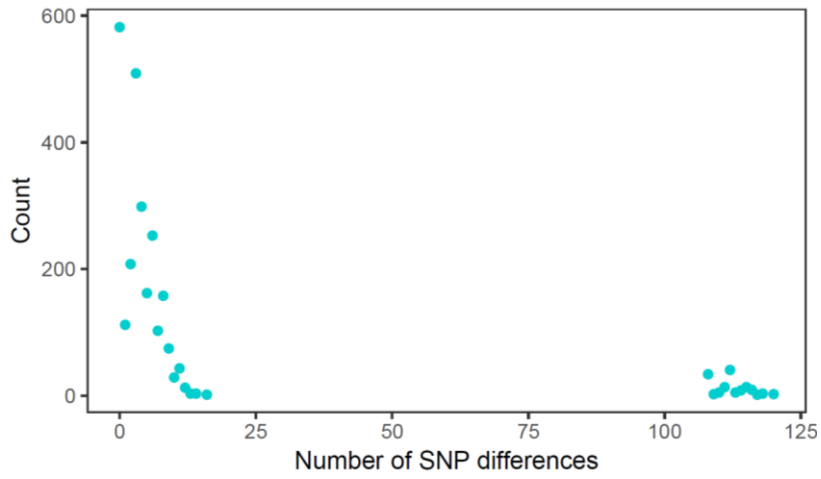
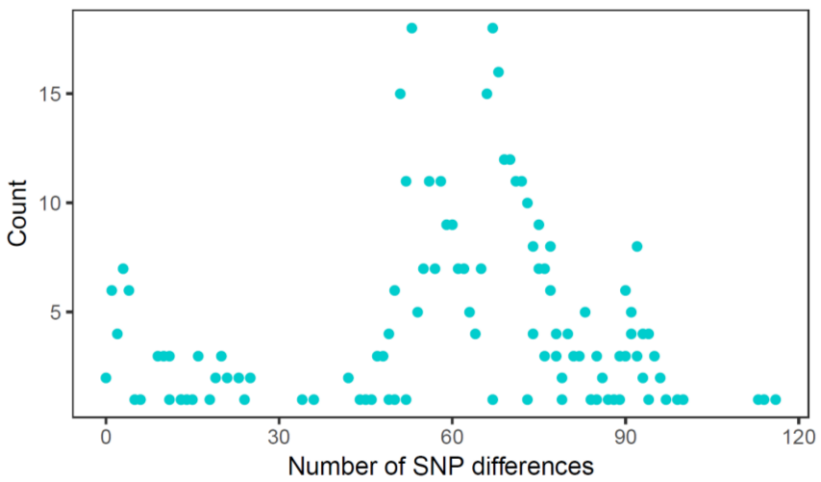


Figure 3. DLST 1-18 maximum likelihood. Non-outbreak isolates belonging to Patient 1 are highlighted in grey, and clustered apart from the remaining isolates. Isolates from Patient 4 and 23 are highlighted in pink and blue, respectively.

(A)



(B)



(C)

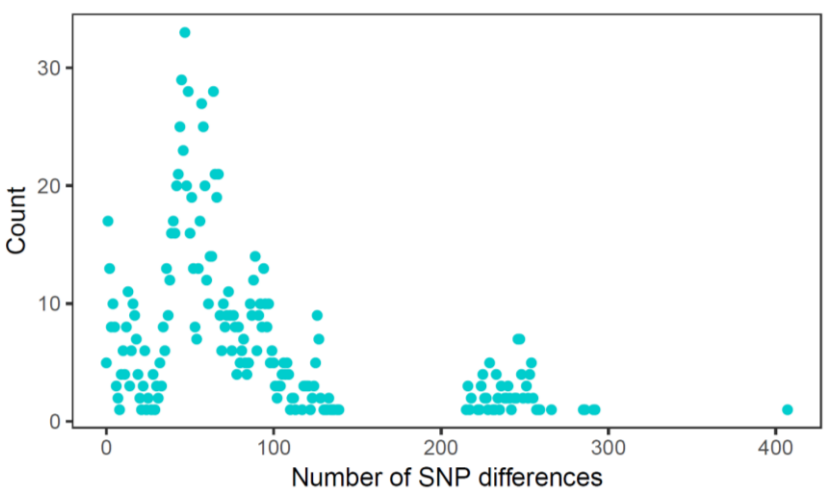


Figure 4. Frequency of number of SNP differences for **(A)** DLST 1-18, **(B)** DLST 1-21, and **(C)** DLST 6-7.

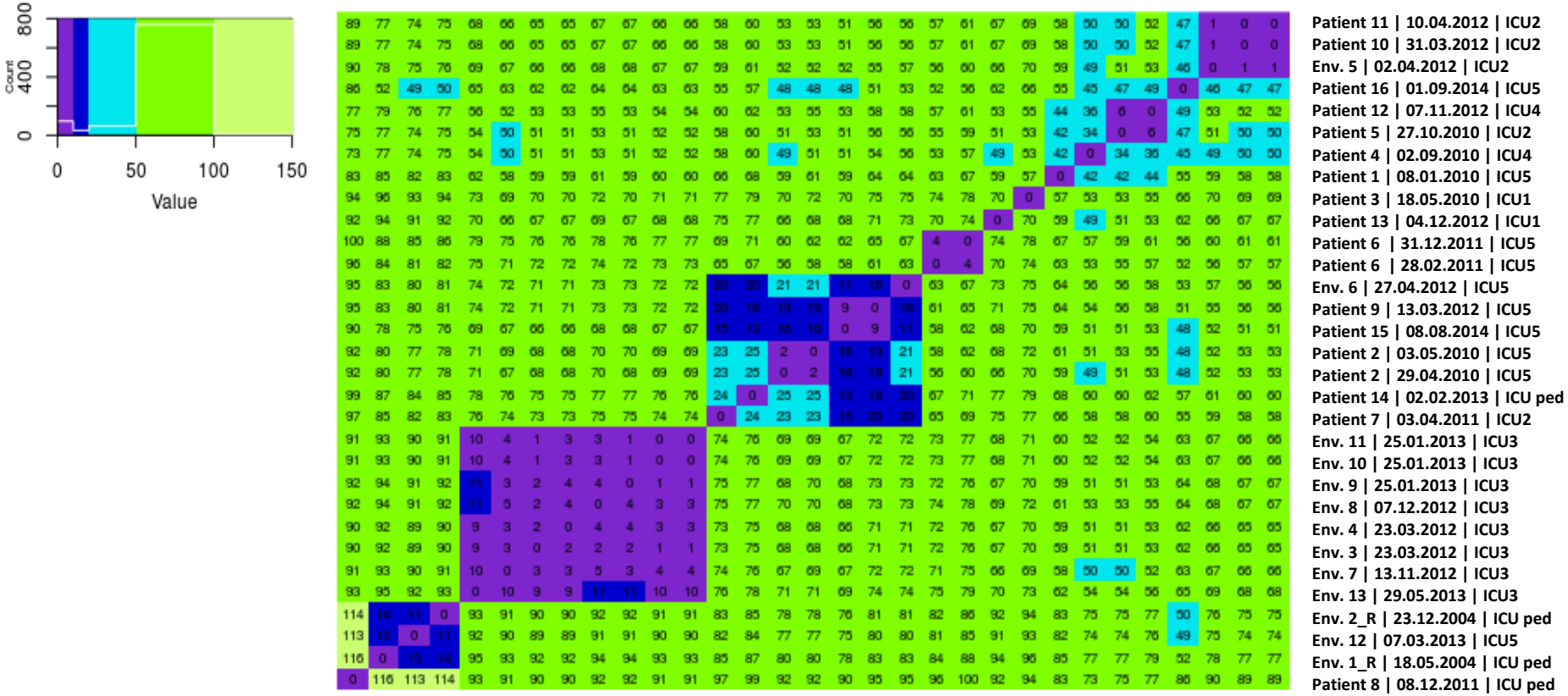


Figure 5. DLST 1-21 color heatmap showing pairwise genomic distances. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Patient 8 (first isolate) to Patient 11 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

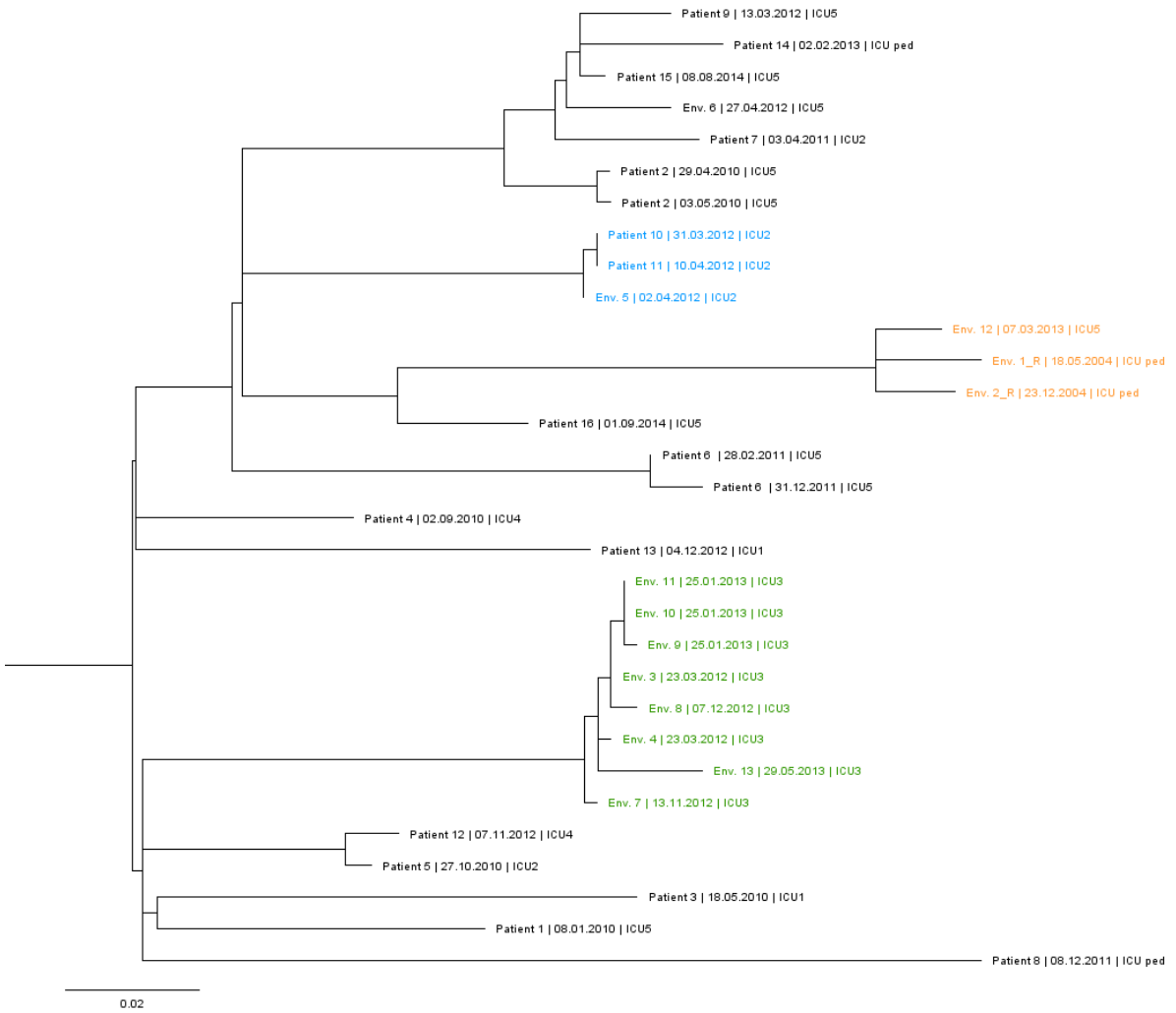


Figure 6. DLST 1-21 maximum likelihood tree. Three environmental isolates retrieved between 2004 and 2013 are highlighted in orange; isolates from two patients and an environmental sample collected from ICU2 are highlighted in blue; subclade of environmental isolates from the burn unit are highlighted in green.

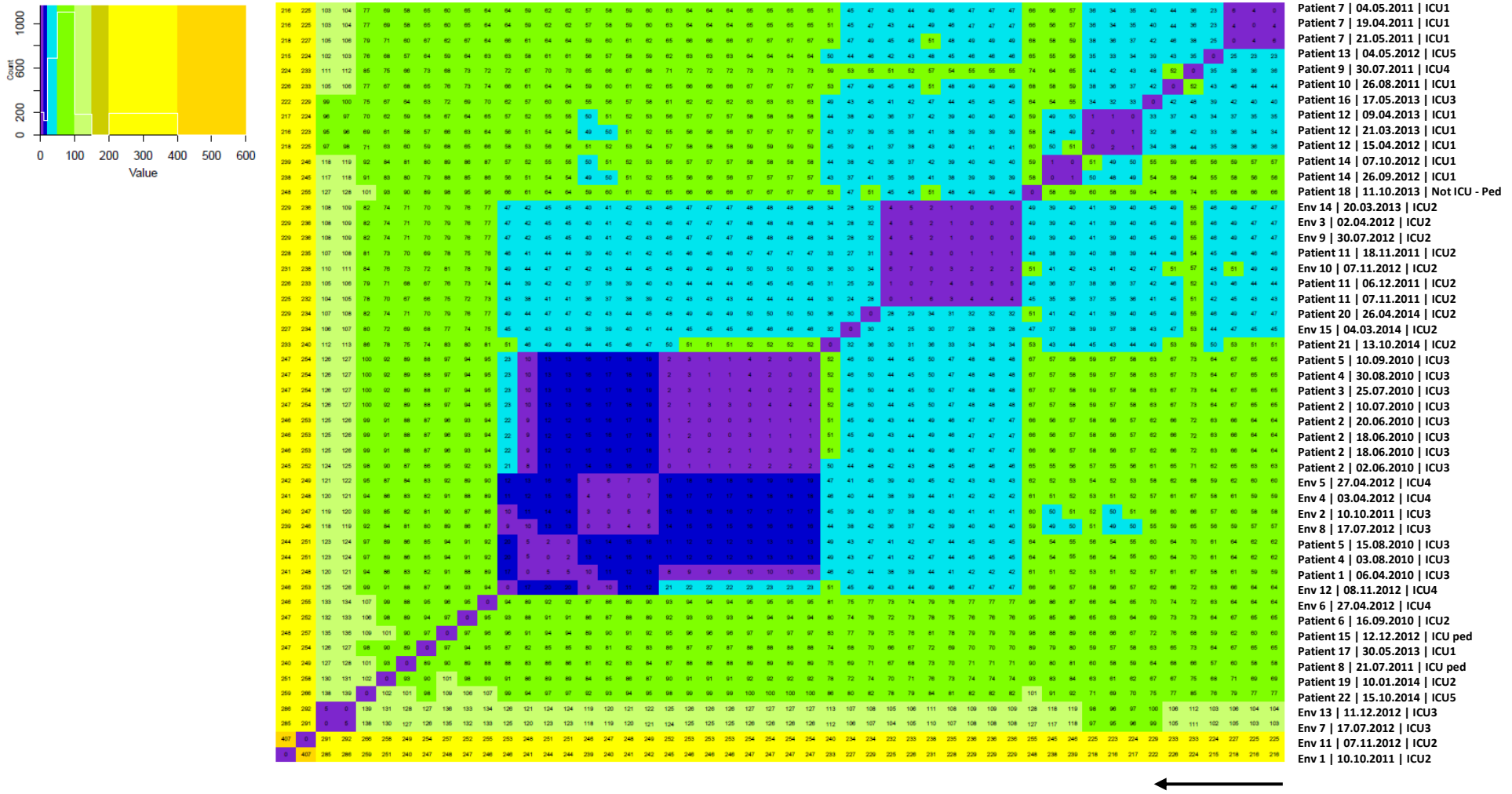


Figure 7. DLST 6-7 color heatmap showing pairwise genomic distances. Number of SNP differences between pairs of isolates are displayed in each square. Each line corresponds to an isolate. Isolate's identification on the columns from left to right: Env. 1 (first isolate) to Patient 7 (last isolate). Different colors represent different SNP differences' limits:10, 20, 50, 100, and 150. The frequency of each number of SNP differences is pictured by a white line on the color legend plot.

Figure 8. DLST 6-7 maximum likelihood tree. A subclade of isolates from the burn unit suspected to be epidemiologically linked are highlighted in green. Another subclade, in blue, is composed of isolates from Patient 11 and environmental isolates retrieved from ICU2. An example of isolates belonging to the same patient, Patient 12, is highlighted in green. Two long branches belonging to Env. 1 and Env. 11 are highlighted in red.

5. References

1. Basset P, Blanc DS. 2014. Fast and simple epidemiological typing of Pseudomonas aeruginosa using the double-locus sequence typing (DLST) method. Eur J Clin Microbiol Infect Dis 33:927-32.
2. Pollack M, Koles NL, Preston MJ, Brown BJ, Pier GB. 1995. Functional properties of isotype-switched immunoglobulin M (IgM) and IgG monoclonal antibodies to Pseudomonas aeruginosa lipopolysaccharide. Infect Immun 63:4481-8.
3. Morrison AJ, Jr., Wenzel RP. 1984. Epidemiology of infections due to Pseudomonas aeruginosa. Rev Infect Dis 6 Suppl 3:S627-42.
4. Blanc DS, Francioli P, Zanetti G. 2007. Molecular Epidemiology of Pseudomonas aeruginosa in the Intensive Care Units - A Review. Open Microbiol J 1:8-11.
5. Lyczak JB, Cannon CL, Pier GB. 2000. Establishment of Pseudomonas aeruginosa infection: lessons from a versatile opportunist. Microbes Infect 2:1051-60.
6. Gellatly SL, Hancock RE. 2013. Pseudomonas aeruginosa: new insights into pathogenesis and host defenses. Pathog Dis 67:159-73.
7. Lister PD. 2002. Chromosomally-encoded resistance mechanisms of Pseudomonas aeruginosa: therapeutic implications. Am J Pharmacogenomics 2:235-43.
8. Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R, McLaughlin SF, Peckham HE, Omberg L, Bormann Chung CA, Stanley S, Pearlstein K, Levandowsky E, Acevedo-Acevedo S, Auton A, Keinan A, Acuna-Alonzo V, Barquera-Lozano R, Canizales-Quinteros S, Eng C, Burchard EG, Russell A, Reynolds A, Clark AG, Reese MG, Lincoln SE, Butte AJ, De La Vega FM, Bustamante CD. 2012. Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am J Hum Genet 91:660-71.
9. Maatallah M, Cheriaa J, Backhrouf A, Iversen A, Grundmann H, Do T, Lanotte P, Mastouri M, Elghmati MS, Rojo F, Mejdi S, Giske CG. 2011. Population structure of Pseudomonas aeruginosa from five Mediterranean countries: evidence for frequent recombination and epidemic occurrence of CC235. PLoS One 6:e25617.
10. Pirnay JP, Bilocq F, Pot B, Cornelis P, Zizi M, Van Eldere J, Deschaght P, Vaneechoutte M, Jennes S, Pitt T, De Vos D. 2009. Pseudomonas aeruginosa population structure revisited. PLoS One 4:e7740.
11. Blanc DS. 2004. The use of molecular typing for epidemiological surveillance and investigation of endemic nosocomial infections. Infect Genet Evol 4:193-7.

12. Gautom RK. 1997. Rapid pulsed-field gel electrophoresis protocol for typing of Escherichia coli O157:H7 and other gram-negative organisms in 1 day. J Clin Microbiol 35:2977-80.
13. Salimi J. 2009. On the management of mycotic femoral pseudoaneurysms in intravenous drug abusers. Ann Vasc Surg 23:824.
14. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233-9.
15. Botes J, Williamson G, Sinickas V, Gurtler V. 2003. Genomic typing of Pseudomonas aeruginosa isolates by comparison of Riboprinting and PFGE: correlation of experimental results with those predicted from the complete genome sequence of isolate PAO1. J Microbiol Methods 55:231-40.
16. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Kohler T, van Delden C, Weinel C, Slickers P, Tummler B. 2007. Population structure of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 104:8101-6.
17. Basset P, Hammer NB, Kuhn G, Vogel V, Sakwinska O, Blanc DS. 2009. Staphylococcus aureus clfB and spa alleles of the repeat regions are segregated into major phylogenetic lineages. Infect Genet Evol 9:941-7.
18. Basset P, Senn L, Prod'hom G, Bille J, Francioli P, Zanetti G, Blanc DS. 2010. Usefulness of double locus sequence typing (DLST) for regional and international epidemiological surveillance of methicilin-resistant Staphylococcus aureus. Clin Microbiol Infect 16:1289-96.
19. Sakwinska O, Blanc DS, Lazor-Blanchet C, Moreillon M, Giddey M, Moreillon P. 2010. Ecological temporal stability of Staphylococcus aureus nasal carriage. J Clin Microbiol 48:2724-8.
20. Dettman JR, Rodrigue N, Aaron SD, Kassen R. 2013. Evolutionary genomics of epidemic and nonepidemic strains of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 110:21065-70.
21. Jelsbak L, Johansen HK, Frost AL, Thogersen R, Thomsen LE, Ciofu O, Yang L, Haagensen JA, Hoiby N, Molin S. 2007. Molecular epidemiology and dynamics of Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients. Infect Immun 75:2214-24.
22. Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, Pallen MJ, Ala'Aldeen DA. 2013. Epidemiological investigation of Pseudomonas aeruginosa isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. Euro Surveill 18.

23. Tissot F, Blanc DS, Basset P, Zanetti G, Berger MM, Que YA, Eggimann P, Senn L. 2016. New genotyping method discovers sustained nosocomial Pseudomonas aeruginosa outbreak in an intensive care burn unit. J Hosp Infect 94:2-7.
24. Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221-4.
25. Cholley P, Stojanov M, Hocquet D, Thouverez M, Bertrand X, Blanc DS. 2015. Comparison of double-locus sequence typing (DLST) and multilocus sequence typing (MLST) for the investigation of Pseudomonas aeruginosa populations. Diagn Microbiol Infect Dis 82:274-7.
26. Kanamori H, Weber DJ, Rutala WA. 2016. Healthcare Outbreaks Associated With a Water Reservoir and Infection Prevention Strategies. Clin Infect Dis 62:1423-35.
27. Cramer N, Wiehlmann L, Ciofu O, Tamm S, Hoiby N, Tummler B. 2012. Molecular epidemiology of chronic Pseudomonas aeruginosa airway infections in cystic fibrosis. PLoS One 7:e50731.
28. Kuck P, Mayer C, Wagele JW, Misof B. 2012. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. PLoS One 7:e36593.
29. Jolivet-Gougeon A, Kovacs B, Le Gall-David S, Le Bars H, Bousarghin L, Bonnaure-Mallet M, Lobel B, Guille F, Soussy CJ, Tenke P. 2011. Bacterial hypermutation: clinical implications. J Med Microbiol 60:563-73.
30. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 52:1501-10.

Supplementary materials

Supplementary data 1

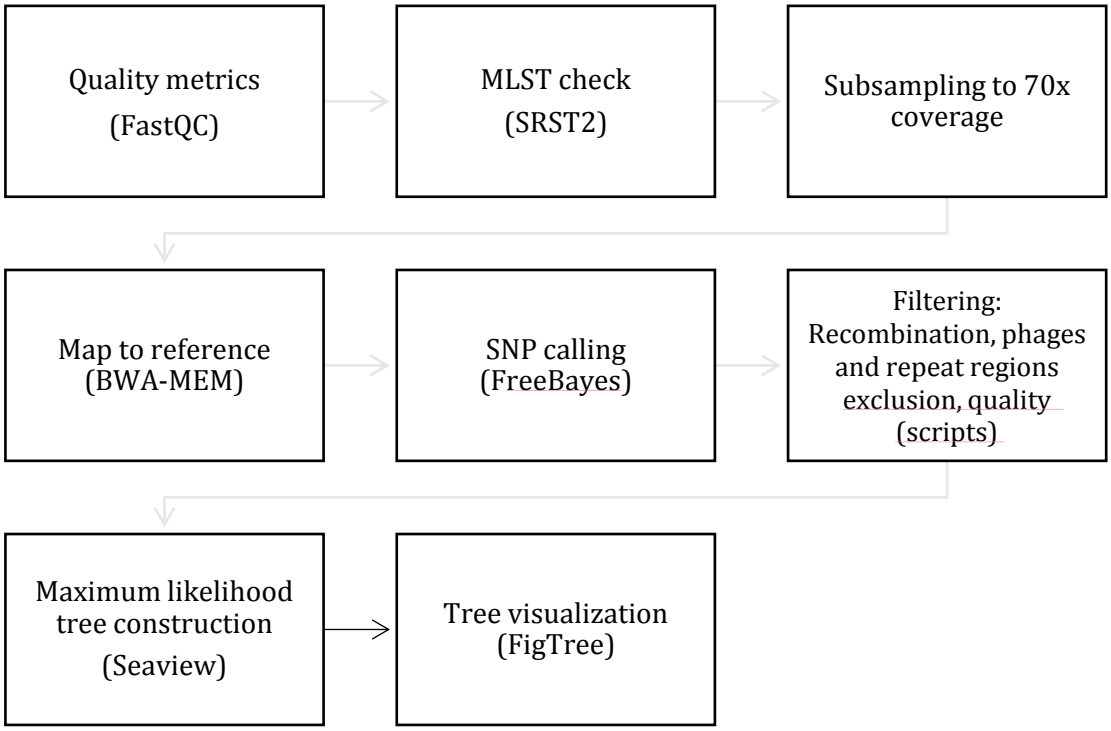


Figure 1. Schematic representation of the different steps included in the adapted methodology.

A first step of subsampling the number of raw reads to reach the lower read depth observed (70x) was done to provide comparable accuracy in the posterior analysis, as well as to reduce mapping time. The subsampled reads were then mapped against their respective complete reference genome with BWA-MEM. Variant calling was performed with FreeBayes with a minimum mapping quality of 60 and a minimum proportion for variant evidence of 0.9. A series of other in-house scripts were applied to the variant call format (VCF) file (lists each position where a SNP is detected along with several characteristics associated with this SNP, such as the nucleotide change, quality value, or the applied filtering) acquired after SNP calling with FreeBayes An in-house script was used for identification of recombination regions: it determines a threshold for SNP density according to the data being analysed and lists the regions of high SNP density to

be masked above this threshold. A probability to remove regions of high SNP density of 0.001 and a window size of 2000 was used for recombination detection. Additionally, an in-house script performed repeat region identification. Putative phages found with PHASTER, along with repeat regions and potential recombination regions were excluded from the genome alignment. The VCF file was then filtered with other in-house scripts applying the following parameter thresholds: minimum quality of base assignement of 100 and a minimum read by allele to report a SNP of 20. A maximum likelihood tree was constructed from the final core SNPs alignment using the PhyML algorithm implemented in Seaview version 4.7 (24). Tree visualization was done with FigTree version 1.4.3.