Application of WGS data for O-specific antigen analysis and in silico serotyping of Pseudomonas aeruginosa isolates

Thrane, Sandra Wingaard; Taylor, Véronique L.; Lund, Ole; Lam, Joseph S.; Jelsbak, Lars

Published in:
Journal of Clinical Microbiology

Link to article, DOI:
10.1128/JCM.00349-16

Publication date:
2016

Document Version
Peer reviewed version

Citation (APA):
Thrane, S. W., Taylor, V. L., Lund, O., Lam, J. S., & Jelsbak, L. (2016). Application of WGS data for O-specific antigen analysis and *in silico* serotyping of *Pseudomonas aeruginosa* isolates. *Journal of Clinical Microbiology*, *54*(7), 1782-1788. https://doi.org/10.1128/JCM.00349-16

JCM Accepted Manuscript Posted Online 20 April 2016
J. Clin. Microbiol. doi:10.1128/JCM.00349-16

Application of WGS data for O-specific antigen analysis and *in silico* serotyping of *Pseudomonas aeruginosa* isolates

Sandra Wingaard Thrane[1], Véronique L. Taylor[2], Ole Lund[3], Joseph S. Lam[2] and Lars Jelsbak[1]#

[1]Technical University of Denmark, Department of Systems Biology, Kgs. Lyngby, Denmark

[2]University of Guelph, Department of Molecular and Cellular Biology, Guelph, Canada

[3]Technical University of Denmark, Center for Biological Sequence Analysis, Department of Systems Biology, Kgs. Lyngby, Denmark

Running title: *In silico* serotyping of *P. aeruginosa*

#Address correspondence to Lars Jelsbak, lj@bio.dtu.dk

Abstract

Accurate typing methods are required for efficient infection control. The emergence of whole genome sequencing (WGS) technologies has enabled the development of genomics-based methods applicable for routine typing and surveillance of bacterial pathogens. In this study, we developed the *Pseudomonas aeruginosa* serotyper (PAst) program, which enabled *in silico* serotyping of *P. aeruginosa* isolates using WGS data. PAst has been made publically available as a web-service, and aptly facilitate high-throughput serotyping analysis. The program overcomes critical issues such as the loss of *in vitro* typeability often associated with *P. aeruginosa* isolates from chronic infections, and quickly determines the serogroup of an isolate based on the sequence of the O-specific antigen (OSA) gene cluster. Here, PAst analysis of 1649 genomes resulted in successful serogroup assignments in 99.27% of the cases. This frequency is rarely achievable by conventional serotyping methods. The limited number of non-typeable isolates found using PAst was the result of either complete absence of OSA genes in the genomes or the artifact of genomic misassembly. With PAst, *P. aeruginosa* serotype data can be obtained from WGS information alone. PAst is a highly efficient alternative to conventional serotyping methods in relation to outbreak surveillance of serotype O12 and other high-risk clones, while maintaining backward compatibility to historical serotype data.

Introduction

Pseudomonas aeruginosa is a Gram-negative opportunistic pathogen and a major cause of mortality and morbidity among hospitalized and compromised patients including those with cystic fibrosis (CF). *P. aeruginosa* is well known for its ability to cause chronic and extensively drug resistant infections (1). The outer membrane lipopolysaccharide (LPS) layer is a major virulence factor of *P. aeruginosa* (2). LPS has been linked to antibiotic resistance and immune evasion. Furthermore, LPS is one of the receptors that determines susceptibility of the bacterium to bacteriophages and pyocins (2–4). Our ability to control *P. aeruginosa* infections depends on the availability of accurate typing methods. Previously, serotyping was a benchmark typing method for *P. aeruginosa*. In the 1980's the International Antigenic Typing Scheme (IATS) was established to classify the species *P. aeruginosa* into 20 serotypes (O1-O20) (5–7). Today, serotyping is infrequently used in the clinic for typing purposes, mainly because of the time consuming protocol, the need for a continuous supply of serotype-specific antisera, and a high prevalence of polyagglutinating or non-typeable isolates.

The loss of *P. aeruginosa* typeability has been known for decades, and has often been linked to bacteria isolated from chronic infections, where typeability is lost over time during the course of infection (8, 9). A study performed by Pirnay *et al* (10) showed that 65% of all *P. aeruginosa* isolates examined were either non- or multi-typeable and therefore assigning a particular serotype to these strains would be difficult. The occurrence of these non- or multi-typeable isolates was higher when evaluating isolates sampled exclusively from CF infections (10). Multi-typeability has been associated with poor prognosis for CF patients, and is a trait of persistent or chronic infection. This correlates with the observation that *P. aeruginosa* isolates from chronic CF infections are initially resistant to human serum but evolve to becoming serum sensitive over time. This is likely due to the loss of production of O-antigen, which protects the bacterial cell from the human serum (8). The mechanism underlying loss of typeability over time is not fully understood, but is most likely due to modifications of LPS structures over extended periods of bacteria-host interactions as a means to improve fitness in the host and to evade host immune system, bacteriophages and antibiotic therapy.

The knowledge concerning the serotype of an isolate is important for monitoring outbreaks and for understanding the structures of the LPS expressed on the surface of these bacteria. O11 and O12 are more predominant than other serotypes in the clinic, and intriguingly, these serotypes have been associated with multi-drug resistance (MDR) (10–13). This implies that these particular LPS structures improve fitness within the hosts and the hospital environments in ways that we currently do not understand. Specifically for the O12 serotype, it has been shown that horizontal gene transfer of LPS genes has resulted in MDR isolates and the switching of a certain serotype to O12 (14). To continuously monitor LPS structure and evolution, serotyping can help to improve our understanding of the isolates that successfully infect patients. The continued collection of these data will also enable retrospective population analysis, as serotype has been recorded for decades also prior to the emergence of other DNA-based typing methods such as MLST and PCR.

P. aeruginosa LPS is comprised of three domains: lipid A, core oligosaccharide, and O-antigen (2). Most *P. aeruginosa* isolates produce two forms of O-antigen simultaneously: common polysaccharide antigen (CPA) and O-specific antigen (OSA). While CPA is relatively conserved, OSA is variable and defines the serotype of an isolate (2, 15). OSA is encoded in a gene cluster varying in size from just under 15 kb to over 25 kb. The OSA gene cluster is flanked by the genes *ihfB*/*himD* and *wbpM.* The 20 serotypes harbor 11 distinct OSA gene clusters, each with a high number of unique genes (16). With the emergence of whole genome sequencing (WGS) methods it is now possible to assign an isolate into one of 11 serogroups based on the sequence and structure of the OSA gene cluster (11, 14, 17).

The present study presents a program that our group has developed for fast and reliable *in silico* serotyping of *P. aeruginosa* isolates using WGS data – the *Pseudomonas aeruginosa* serotyper (PAst). The program has been made publically available as a web-service, and can enable high throughput serotyping analysis based on analysis of the OSA gene cluster. Using PAst, issues with typeability of clinical isolates can be overcome, and serotyping can be performed in a rapid and cost-effective way in the clinic as whole genome sequencing of isolates become accessible.

Materials and Methods

PAst verification and isolates included in the study

To evaluate the efficiency of the *in silico* serotyping using PAst, all available *P. aeruginosa* genomes were acquired and analyzed. These *P. aeruginosa* genomes were downloaded from NCBI and included 1120 genome assemblies (Supplementary Table 1, extracted 18.08.2015). An exclusively CF-related *P. aeruginosa* dataset was constructed, due mainly to the documented high level of non-typeability in persistent infecting clones. The isolates described by Marvig *et al.* 2015 (475 genomes) (18) were used as the initial dataset. These were assembled using SPAdes (19) prior to analysis. Additional CF isolates were recovered by searching for *P. aeruginosa* genome assemblies related to CF in PATRIC (54 genomes) (20). It was verified that frequently observed CF-specific strains such as DK2 and LES were part of the dataset. The final dataset included 529 CF-related *P. aeruginosa* genome assemblies. *In silico* serotyping of both datasets was performed using PAst in order to evaluate typeability of the program. Non-typeable isolates (i.e., isolates in which %coverage of reference OSA was < 95%) were manually examined for either biological or technical explanations of the lack of typeability.

PAst specifications

The PAst program is developed using the programming language Perl for *in silico* serotyping of *P. aeruginosa* isolates using WGS data. It is based on a BLASTn analysis of the assembled input genome, against an OSA cluster database. OSA clusters with > 95% coverage in the query genome represents a positive hit for a serogroup. Since *P. aeruginosa* isolates have been described which either harbor multiple OSA clusters or no clusters at all, the program accommodates multi-, mono- and non-typeability based on analysis of the number of positive OSA hits and coverage (Figure 1). Compared to other studies (11, 14, 17) PAst optimizes *in silico* serotyping further by distinguishing members of the O2 serogroup through identification of the acquired phage-related wzy_{β} within serotypes O2 and O16 (21, 22). This enables typing into 12 serogroups as opposed to the 11 described by Raymond *et al.* (16).

Together with a summary of the best hit(s) from the analysis and the BLAST report, the user receives a multi fasta file containing the sequence(s) of the OSA cluster from the analyzed isolate for use in future analysis.

The *P. aeruginosa* OSA cluster database

The database was constructed using the WGS data of the 20 *P. aeruginosa* IATS serotype reference isolates (14). The genomes were assembled using SPAdes (19) and the OSA clusters extracted via identification of the *ihfB/himD* gene flanking the cluster upstream and the *wbpM* gene flanking the cluster downstream. The clusters were aligned within their serotypes, described by Raymond *et al.* 2002 and their shared structure confirmed (16). A representative cluster of each serotype was selected for the database (Table 1). Also included in the database was the wzy_{β} gene for distinguishing the O2 and O5 serotypes, as the two serogroups share OSA cluster organization, but only the O2 and O16 serotype harbor the wzy_{β} gene present on a prophage.

***In silico* serotyping of *P. aeruginosa* isolates using PAst**

PAst has been implemented as a simple and user-friendly web-tool available on the Center for Genomic Epidemiology (CGE) service platform (https://cge.cbs.dtu.dk/services/PAst-1.0/). The tool accommodates raw reads, draft assemblies (contigs or scaffolds) and complete genomes from all WGS platforms. Raw read data are processed and assembled as previously described for other CGE tools (23). Following analysis of the input data, the web-tool outputs the predicted serogroup of the query genome, the %coverage of the reference OSA cluster, as well as the OSA cluster sequence in multi fasta format, for the user to continue exploring the OSA genes (Fig. 1). If multiple positive hits are found (multi-typeability), all the identified OSA clusters are written for the user (Fig. 1). In the case of a non-typeable query genome (where no OSA cluster has >95% coverage) the best hit identified is written for the user together with the sequence of this hit (Fig. 1).

For batch analysis of larger datasets (only applicable for assembled genomes) the PAst Perl program has been made available on Github: https://github.com/Sandramses/PAst

Results

The PAst web server tool identifies and analyzes the nucleotide sequence of the O-specific antigen (OSA) gene cluster within the provided genomes and place them into one of twelve serogroups defined in Table 1. These serogroups are defined by sequence similarities between the 20 IATS serotypes (16) as well as absence/presence of the discriminatory wzy_{β} gene (21, 22) and are as such different from previously groupings of serotypes on the basis of *in vitro* serotyping data (11, 14, 17). All serogroups contained three or less of the 20 IATS serotypes (Table 1).

More than 97% of the *P. aeruginosa* dataset is typeable using PAst

To evaluate the typeability efficiency of PAst all *P. aeruginosa* genome assemblies available in NCBI (1120 genomes on date of extraction) were analyzed. A total of 97.68% (1094) of the 1120 genomes were typed unambiguously to a single serogroup by PAst (Fig. 2). This means that each genome assembly had a single BLAST hit of >95% OSA coverage to one sequence in our reference OSA database (Fig. 2). No isolates were found to be multi-typeable and 2.32% (26 genomes) of the 1120 genomes were found to be non-typeable (Fig. 2). In these cases, no significant BLAST hit of >95% OSA coverage to one of the sequence in the reference OSA database was identified. PAst correctly determined the serogroup of the 20 IATS strains as well as PAO1 (serotype O5), PA14 (serotype O10), and PAK (serotype O6).

The analysis showed that all serogroups were represented in the 1120 genomes (Fig. 2). Four of the 12 serogroups represented 70% of the genomes analyzed; these were O3, O6, O11 and O12 (Fig. 2). The smallest serogroup was O13, which contained only four genomes. We note that the same clone type could be present multiple times in the dataset, and that a substantial sampling bias would therefore be expected. The distribution of serotypes in our analysis thus describes what has been chosen for sequencing and does not necessarily match the distribution of serotypes in the actual *P. aeruginosa* population. This does not affect the high confidence of PAst, as it shows that un-ambiguous typing of multiple isolates from the same lineage is possible.

PAst overcomes non-typeability issues from *in vitro* typing of CF lineages

P. aeruginosa isolates from CF infections are often non-typeable with conventional serotyping assays. To explore if our genomics-based method could enable acquisition of serotype information in such isolates, we analyzed 529 genome assemblies of *P. aeruginosa* isolates sampled from CF infections. This dataset contained multiple examples of isolates of the same lineage that had been sampled during the course of infection. This enabled us to investigate whether *in silico* typeability might be lost over time as has frequently been observed for *in vitro* serotyping of isolates from chronic CF infections. Interestingly, 99.81% of the genomes in the CF-specific dataset could be typed to single serogroups. More importantly, no multi-typeable isolates were observed and only one isolate was deemed non-typeable (Fig. 3). All serogroups were represented in the dataset, except for O12. The absence of O12 serotypes among CF isolates has previously been reported (10). Serotypes O1, O6 and O7/O8 represented ~65% of the CF-specific dataset and the smallest representation of serotypes was the O9 serogroups with only two isolates from these samples (Fig. 3).

Well-known transmissible CF-specific clone types such as *P. aeruginosa* DK1 (24), DK2 (25), and LES (26) are represented in the dataset due to multiple isolates being sampled from various patients over several decades. Using our PAst tool, the typing problems documented from *in vitro* typing of such lineages were not observed, and the DK1, DK2 and LES isolates were consistently *in silico* serotyped with PAst. DK1 and DK2 were found to belong to the O3 serogroup, while the LES lineage belonged to the O6 serogroup.

Complete loss of O-specific antigen defining genes is a rare event

Out of two WGS-based datasets (n = 1649) that were *in silico* typed with PAst, our results yielded a total of 27 non-typeable isolates. The lack of typeability in these 27 genome assemblies was further investigated to resolve whether non-typeability in these cases was due to technical or biological reasons. We found that the %OSA coverage of the non-typeable isolates ranged from a minimum of 1.91% to a maximum 93.96% OSA coverage (Supplementary Table 2). Of the 27 isolates classified as non-typeable, thirteen were found to have OSA coverage of 0-20%, whereas seven isolates had OSA coverage of 80-95% (Fig. 4). The best hit (serogroup) for each of the non-typeable isolates was then examined to evaluate if certain serogroups were more prone to be problematic in the PAst analysis and why. The 27 isolates were found to distribute across 6 serogroups (O1, O2, O6, O7, O11 and O13), while 15/27 isolates showed a best hit to be typed as the O11 serogroup (Fig. 4).

The group of non-typeable isolates with a best hit to the O11 serogroup were analyzed separately to identify the reason for the lack of typeability. Of the 15 O11 serogroup isolates, nine had an OSA coverage of 14.94-15.84% (Supplementary Table 2); these corresponded to the presence of only the two flanking genes *himD/ihfB* and *wbpM*. This observation shows that a best hit of a non-typeable isolate to the O11 OSA cluster with a coverage of ~15% is the result of a complete absence of an OSA cluster but the presence of the flanking genes. Two other isolates had an OSA coverage of <2%, and corresponded to the absence of the entire OSA cluster as well as the flanking genes (Supplementary table 2). In summary, a total of 11 of the 27 non-typeable isolates (or 11 of 1649 isolates analyzed in total) were non-typeable due to a lack of the OSA cluster sequences.

Genome mis-assembly accounts for false non-typeability

Since the seven non-typeable isolates with the highest OSA coverage (80-95%) in Figure 4 were all candidates for harboring complete and functioning OSA clusters, we analyzed the cause of non-typeability in this group of isolates. For each of the isolates, we examined whether there were mis-assembly or assembly gaps within the OSA gene cluster; we also looked for the occurrence of insertion sequence (IS) elements, which often cause gaps in *de novo* assembly. Indeed, five of the seven isolates contained assembly gaps within their OSA cluster, which account for the observed lowered OSA coverage (Table 2). The remaining two isolates had no gaps within their OSA sequence (Table 2). However, both of these isolates had a best type hit to the O11 serogroup, which is known to contain OSA sequences of both the O11 and the O17 serotypes (16) (Table 1). Interestingly, the OSA cluster in these two serotypes differ only by the presence of two IS elements and a deletion in the O17 serotype OSA sequence (16). Alignment of the OSA sequence from the two non-typable isolates to the O11 and O17 reference OSA sequences, respectively, contained an O17 OSA gene cluster, which had been misassembled into concatenated O11 serotype OSA clusters because of the O17 IS elements.

Discussion

The serotyping technique has been one of the standard tools for epidemiological studies and infection controls for many decades. The available historical records of *P. aeruginosa* serotypes offer a vast amount of information about *P. aeruginosa* epidemiology and population structures (27–30). Although problems with non-typeable isolates have been described since the implementation of the method, the serotype information is still applicable today for outbreak tracking, strain typing, and studies of LPS structure and evolution. The present study presents a newly developed Web Server tool called PAst, which is user friendly, reliable, and high-throughput for *in silico* serotyping of *P. aeruginosa* isolates.

In contrast to conventional serology-based *in vitro* serotyping, PAst *in silico* serotyping has a very low occurrence of non-typeablility. Of the 1649 analyzed genomes, only 27 non-typeable isolates were detected across two separate *P. aeruginosa* datasets. One dataset represents all available whole genome assemblies of *P. aeruginosa,* while the other specifically represents genomes from CF infections, which are known to contain high occurrences of non-typeability due to adaptability of the bacteria into a biofilm life-style associated with chronicity of the infection (Fig. 1 and 2). Importantly, since the frequency of non-typeability of *in vitro* serotyped *P. aeruginosa* isolates may amount to over 65% (10), analysis with PAst is clearly advantageous and superior compared to conventional *in vitro* serotyping. Importantly, the superiority of the PAst tool as a reliable and fast typing method is consistent with other published tools for *in silico* serotyping (31–35). Similar to both the SerotypeFinder (*in silico* serotyping of *E. coli* (31)), LisSero (*in silico* serotyping of *Listeria monocytogenes* (34, 35)) and SeqSero (*in silico* serotyping of *Salmonella* (32)) PAst resolves the OSA cluster information to the most accurate typing possible as a serogroup representing 1-3 serotypes.

Interestingly, we observed a high level of conservation of the OSA gene cluster within the *P. aeruginosa* genome. In contrast to certain well-documented difficulties in serology-based *in vitro* serotyping, PAst identified complete OSA clusters (with >95% sequence being present) in 99.27% of the analyzed genomes. As such only 12 of the 1649 isolates examined were found to be devoid of the OSA cluster and an additional 8 isolates were found to contain only a partial OSA cluster in their genomes (<80% OSA sequence compared to the reference). These findings indicate that the loss of typeability of *P. aeruginosa* isolates during the course of infection is either due to mutations (rather than larger deletions) or is linked to other parts of the LPS biosynthesis, such as regulatory genes or transport of the structure to the cell surface. A study by Bélanger *et al.* reported that mutation in any of the four *wbp* genes (*wbpO*, *wbpP*, *wbpV* and *wbpM*) in the OSA gene cluster could disrupt the *P. aeruginosa* O6 OSA biosynthesis (36). Furthermore, key genes involved in the OSA assembly and translocation through the Wzx/Wzy-dependent pathway not localized within the OSA cluster, for instance, *waaL*, are essential for O-antigen expression (37, 38). It is possible that more OSA-related genes might be present in the *P. aeruginosa* genomes, which have not been discovered yet. Overall, our study demonstrates that a complete lack of an OSA gene cluster is a rarely observed phenomenon in *P. aeruginosa*.

PAst will enable further investigations of the diversity, evolution and variability of the OSA clusters. For example, the sequence of the cluster is part of the output material from the *in silico* serotyping which can then be readily analyzed for sequence variations to provide new knowledge on the mechanisms behind loss of typeability *in vitro* and *in silico*. Furthermore, PAst will enable systematic analysis of serotype switching by horizontal gene transfer and genetic recombination of the OSA gene cluster among different clone types. This recently described phenomenon has contributed to the evolution of the multi-drug resistant *P. aeruginosa* serotype O12 population that has successfully disseminated across hospitals worldwide (14). It is currently unknown if there are additional cases of such serotype switching by recombination.

The new PAst Web Server tool makes *in silico* serotyping of *P. aeruginosa* using WGS data a fast and reliable method. The use of PAst can play an important role in future surveillance of LPS evolution and possible outbreak detection. With the emergence of rapidly disseminating, high-risk clones of *P. aeruginosa*, such as the O12 ST111 clone, new and reliable typing techniques for improved monitoring and tracking of such outbreaks are becoming increasingly important (13). With the lowered cost of sequencing and the increased focus on WGS of pathogens in clinics and hospital settings, genomics-based tools can assist in designing future treatments and containment of outbreaks.

Acknowledgements

Funding for this study was provided by operating grants from the Villum Foundation to L.J. (VKR023113) and from the Canadian Institutes of Health Research (CIHR) to J.S.L. (MOP-14687). We thank the Center for Genomic Epidemiology (CGE) at the Center for Biological Sequence analysis (CBS) at DTU, especially Johanne Ahrenfeldt and Rosa Allesøre, for expert assistance in setting up the PAst web-service and hosting it on their web-servers. Additional support was provided by the 'A.N. Neergaard og Hustrus' Foundation to L.J., and a travel grant from Knud Højgaards Foundation to S.W.T. V.L.T. was a recipient of a Cystic Fibrosis Canada Doctoral Studentship, Queen Elizabeth II Graduate Scholarships in Science and Technology (QEII-GSST) and J.S.L. holds a Canada Research Chair in Cystic Fibrosis and Microbial Glycobiology.

References

1. **Folkesson A**, **Jelsbak L**, **Yang L**, **Johansen HK**, **Ciofu O**, **Høiby N**, **Molin S**. 2012. Adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis airway: an evolutionary perspective. Nat Rev Microbiol **10**:841–51.

2. **Lam JS**, **Taylor VL**, **Islam ST**, **Hao Y**, **Kocíncová D**. 2011. Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide. Front Microbiol **2**:118.

3. **Köhler T**, **Donner V**, **van Delden C**. 2010. Lipopolysaccharide as shield and receptor for R-pyocin-mediated killing in *Pseudomonas* aeruginosa. J Bacteriol **192**:1921–1928.

4. **Nakayama K**, **Takashima K**, **Ishihara H**, **Shinomiya T**, **Kageyama M**, **Kanaya S**, **Ohnishi M**, **Murata T**, **Mori H**, **Hayashi T**. 2000. The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage. Mol Microbiol **38**:213–231.

5. **Liu P V**, **Wang S**. 1990. Three new major somatic antigens of *Pseudomonas aeruginosa*. J Clin Microbiol **28**:922–925.

6. **Stanislavsky E**, **Lam J**. 1997. *Pseudomonas aeruginosa* antigens as potential vaccines. FEMS Microbiol Rev **21**:243–277.

7. **Liu P V.**, **Matsumoto H**, **Kusama H**, **Bergan T**. 1983. Survey of heat-stable, major somatic antigens of *Pseudomonas aeruginosa*. Int J Syst Bacteriol **33**:256–264.

8. **Penketh A**, **Pitt T**, **Roberts D**, **Hodson M**, **Batten J**. 1983. The relationship of phenotype changes in *Pseudomonas aeruginosa* to the clinical condition of patients with cystic fibrosis. Am Rev Respir Dis **127**:605–608.

9. **Ojeniyi B**. 1994. Polyagglutinable *Pseudomonas aeruginosa* from cystic fibrosis patients - a survey. APMS **102**.

10. **Pirnay J-P**, **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. **Witney AA**, **Gould KA**, **Pope CF**, **Bolt F**, **Stoker NG**, **Cubbon MD**, **Bradley CR**, **Fraise A**, **Breathnach AS**, **Butcher PD**, **Planche TD**, **Hinds J**. 2014. Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of *Pseudomonas aeruginosa*. Clin Microbiol Infect **20**:O609–O618.

12. **Cholley P**, **Thouverez M**, **Hocquet D**, **Van Der Mee-Marquet N**, **Talon D**, **Bertrand X**. 2011. Most multidrug-resistant *Pseudomonas* aeruginosa isolates from hospitals in eastern france belong to a few clonal types. J Clin Microbiol **49**:2578–2583.

13. **Oliver A**, **Mulet X**, **López-causapé C**, **Juan C**. 2015. The increasing threat of *Pseudomonas aeruginosa* high-risk clones. Drug Resist Updat **22**:41–59.

14. **Thrane SW**, **Taylor VL**, **Freschi L**, **Kukavica-ibrulj I**, **Boyle B**, **Laroche J**, **Pirnay J-P**, **Lévesque RC**, **Lam JS**, **Jelsbak L**. 2015. The widespread multidrug-resistant serotype O12 *Pseudomonas aeruginosa* clone emerged through concomitant horizontal transfer of serotype antigen and antibiotic resistance gene clusters. MBio **6**:1–10.

15. **King JD**, **Kocíncová D**, **Westman EL**, **Lam JS**. 2009. Review: Lipopolysaccharide biosynthesis in *Pseudomonas aeruginosa*. Innate Immun **15**:261–312.

16. **Raymond CK**, **Sims EH**, **Kas A**, **Spencer DH**, **Kutyavin T V**, **Ivey RG**, **Zhou Y**, **Kaul R**, **Clendenning JB**, **Olson M V**. 2002. Genetic variation at the O-antigen biosynthetic locus in *Pseudomonas aeruginosa*. J Bacteriol **184**:3614–3622.

17. **Kos VN**, **Déraspe M**, **McLaughlin RE**, **Whiteaker JD**, **Roy PH**, **Alm R a.**, **Corbeil J**, **Gardner H**. 2015. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. Antimicrob Agents Chemother **59**:427–436.

18. **Marvig RL**, **Sommer LM**, **Molin S**, **Johansen HK**. 2015. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. Nat Genet **47**:57–64.

19. **Bankevich A**, **Nurk S**, **Antipov D**, **Gurevich A a**, **Dvorkin M**, **Kulikov AS**, **Lesin VM**, **Nikolenko SI**, **Pham S**, **Prjibelski AD**, **Pyshkin A V**, **Sirotkin A V**, **Vyahhi N**, **Tesler G**, **Alekseyev M a**, **Pevzner P a**. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19**:455–77.

20. **Wattam AR**, **Abraham D**, **Dalay O**, **Disz TL**, **Driscoll T**, **Gabbard JL**, **Gillespie JJ**, **Gough R**, **Hix D**, **Kenyon R**, **Machi D**, **Mao C**, **Nordberg EK**, **Olson R**, **Overbeek R**, **Pusch GD**, **Shukla M**, **Schulman J**, **Stevens RL**, **Sullivan DE**, **Vonstein V**, **Warren A**, **Will R**, **Wilson MJC**, **Yoo HS**, **Zhang C**, **Zhang Y**, **Sobral BW**. 2014. PATRIC , the bacterial bioinformatics database and analysis resource. Nucleic Acids Res **42**:581–591.

21. **Kaluzny K**, **Abeyrathne PD**, **Lam JS**. 2007. Coexistence of two distinct versions of O-antigen polymerase, Wzy-Alpha and Wzy-Beta, in *Pseudomonas aeruginosa* serogroup O2 and their contributions to cell surface diversity. J Bacteriol **189**:4141–4152.

22. **Newton GJ**, **Daniels C**, **Burrows LL**, **Kropinski AM**, **Clarke AJ**, **Lam JS**. 2001. Three-component-mediated serotype conversion in *Pseudomonas aeruginosa* by bacteriophage D3. Mol Microbiol **39**:1237–1247.

23. **Larsen M V.**, **Cosentino S**, **Rasmussen S**, **Friis C**, **Hasman H**, **Marvig RL**, **Jelsbak L**, **Sicheritz-Pontén T**, **Ussery DW**, **Aarestrup FM**, **Lund O**. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol **50**:1355–1361.

24. **Markussen T**, **Marvig L**, **Gómez-lozano M**, **Aanæs K**, **Burleigh AE**, **Høiby N**. 2014. Environmental heterogeneity drives within-host diversification and evolution of *Pseudomonas aeruginosa*. MBio **5**:1–10.

25. **Marvig RL**, **Johansen HK**, **Molin S**, **Jelsbak L**. 2013. Genome analysis of a

transmissible lineage of Pseudomonas aeruginosa reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. PLoS Genet **9**:e1003741.

26. **Jeukens J**, **Boyle B**, **Kukavica-Ibrulj I**, **Ouellet MM**, **Aaron SD**, **Charette SJ**, **Fothergill JL**, **Tucker NP**, **Winstanley C**, **Levesque RC**. 2014. Comparative genomics of isolates of a *Pseudomonas aeruginosa* epidemic strain associated with chronic lung infections of cystic fibrosis patients. PLoS One **9**:e87611.

27. **Lam JS**, **MacDonald LA**, **Kropinski AM**, **Speert DP**. 1988. Characterization of nontypable strains of *Pseudomonas aeruginosa* from cystic fibrosis patients by means of monoclonal antibodies and SDS□polyacrylamide gel electrophoresis. Serodiag Immunother Infect Dis **2**:365–374.

28. **Ojeniyi B**, **Wolz C**, **Doring G**, **Lam JS**, **Rosdahl VT**, **Høiby N**. 1990. Typing of polyagglutinable *Pseudomonas aeruginosa* isolates from cystic fibrosis patients. Acta Pathol Microbiol Immunol Scand **98**:423–431.

29. **Ojeniyi B**, **Lam JS**, **Høiby N**, **Rosdahl VT**. 1989. A comparison of the efficiency in serotyping of *Pseudomonas aeruginosa* from cystic fibrosis patients using monoclonal and polyclonal antibodies. Acta Pathol Microbiol Immunol Scand **97**:631–636.

30. **Speert DP**, **Campbell M**, **Puterman ML**, **Govan J**, **Doherty C**, **Høiby N**, **Ojeniyi B**, **Lam JS**, **Ogle JW**, **Johnson Z**, **Paranchych W**, **Sastry PA**, **Pitt TL**, **Lawrence L**. 1994. A multicenter comparison of methods for typing strains of *Pseudomonas aeruginosa* predominantly from patients with cystic fibrosis. J Infect Dis **169**:134–142.

31. **Joensen KG**, **Tetzschner AMM**, **Iguchi A**, **Aarestrup FM**, **Scheutz F**. 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* using whole genome sequencing (WGS) data. J Clin Microbiol JCM.00008–15.

32. **Zhang S**, **Yin Y**, **Jones MB**, **Zhang Z**, **Deatherage Kaiser BL**, **Dinsmore B a.**, **Fitzgerald C**, **Fields PI**, **Deng X**. 2015. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol **53**:JCM.00323–15.

33. **Yoshida CE**, **Kruczkiewicz P**, **Laing CR**, **Lingohr EJ**, **Victor P**. 2016. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLoS One **11**:e0147101.

34. **Kwong JC**, **Mercoulia K**, **Tomita T**, **Easton M**, **Li HY**, **Bulach DM**, **Stinear TP**, **Seemann T**, **Howden BP**. 2015. Prospective whole genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol **54**:333–342.

35. **Doumith M**, **Buchrieser C**, **Glaser P**, **Jacquet C**, **Martin P**. 2004. Differentiation of the Major *Listeria monocytogenes* serovars by multiplex PCR. J Clin Microbiol **42**:3819–3822.

36. **Bélanger M**, **Burrows LL**, **Lam JS**. 1999. Functional analysis of genes

responsible for the synthesis of the B-band O-antigen of *Pseudomonas aeruginosa* serotype O6 lipopolysaccharide. Microbiology **145**:3505–3521.

37. **Berry MC**, **Mcghee GC**, **Zhao Y**, **Sundin GW**. 2008. Effect of a *waaL* mutation on lipopolysaccharide composition, oxidative stress survival, and virulence in *Erwinia amylovora*. FEMS Microbiol Lett **291**:80–87.

38. **Abeyrathne PD**, **Daniels C**, **Poon KKH**, **Matewish MJ**, **Lam JS**. 2005. Functional characterization of WaaL, a ligase associated with linking O-antigen polysaccharide to the core of *Pseudomonas aeruginosa* lipopolysaccharide. J Bacteriol **187**:3002–3012.

Figure legends

FIG 1 Workflow illustrating the *in silico* serotyping of the *Pseudomonas aeruginosa* serotyper (PAst).

FIG 2 The distribution of the different serogroups (in %) identified via *in silico* serotyping of the *P. aeruginosa* dataset using PAst. The analysis is based on all available *P. aeruginosa* genomes assemblies (n = 1120).

FIG 3 The distribution of the different serogroups (in %) identified via *in silico* serotyping of CF specific *P. aeruginosa* isolates (n = 529) using PAst.

FIG 4 Best-hit serotype distribution of the 27 non-typeable isolates as a function of the OSA coverage.

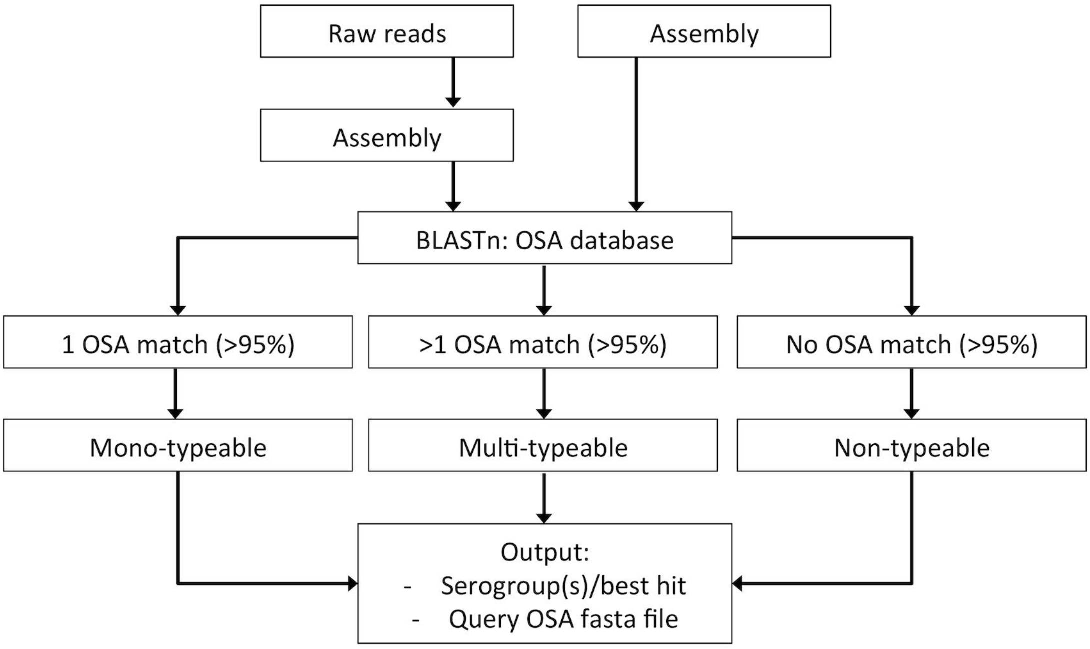
Tables

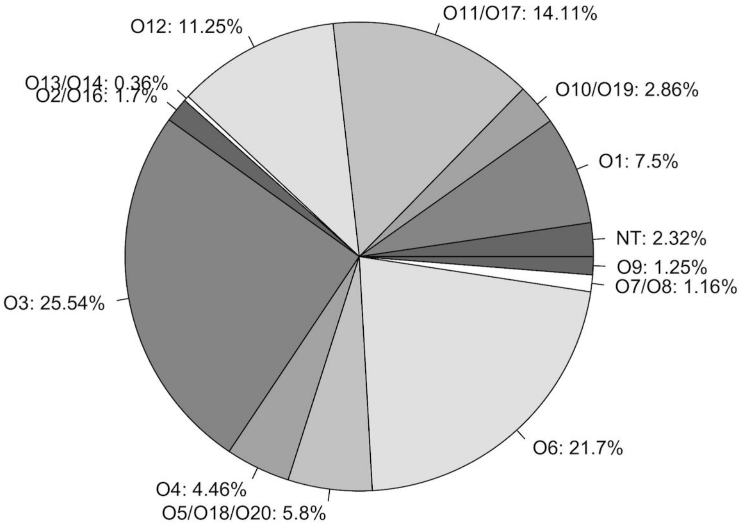
TABLE 1 Serogroup definition in the PAst OSA database.

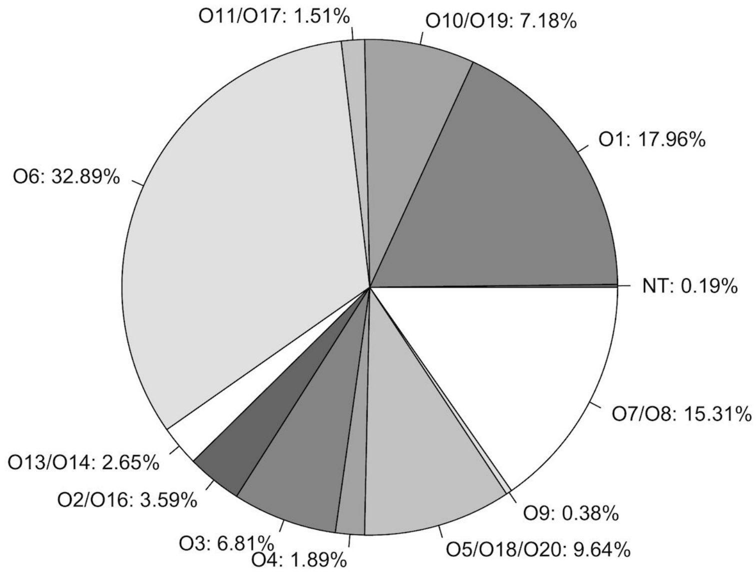
Serogroup	Reference OSA cluster	Ref. gene	Size (bp)	Serotypes within serogroup
O1	O1		18.368	O1
O2	O2	*wzy_{β}*	23.303	O2, O16
O3	O3		20.210	O3, O15
O5	O2		23.303	O5, O18, O20
O4	O4		15.279	O4
O6	O6		15.649	O6
O7	O7		19.617	O7, O8
O9	O9		17.263	O9
O10	O10		17.635	O10, O19
O11	O11		13.868	O11, O17
O12	O12		25.864	O12
O13	O13		14.316	O13, O14

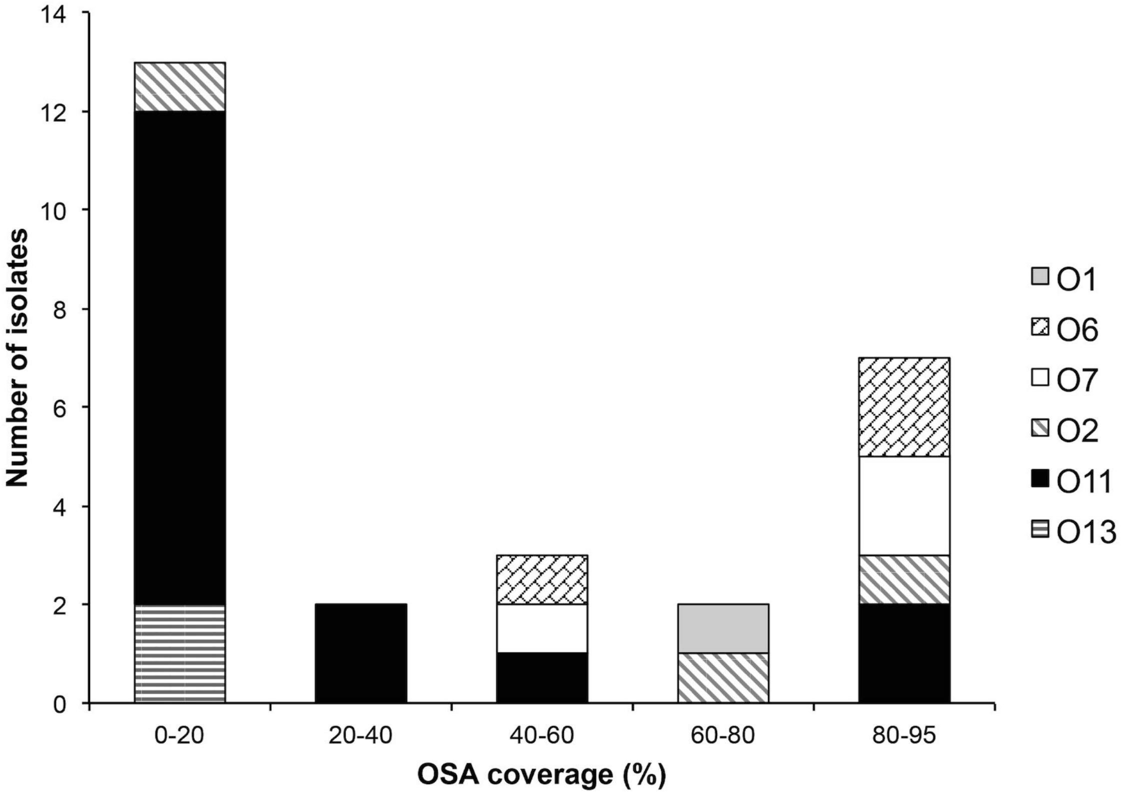
TABLE 2 Non-typeable *P. aeruginosa* isolates with %OSA coverage of 80-95% with specification of assemblies.

Strain	Size (Mb)	Scaffolds	%GC	Best hit	%OSA	*wbpM*	*himD*	Gap
P. aeruginosa E2	635.733	196	66.4	O7	83.31	+	+	+
P. aeruginosa IGB83	648.065	249	66.4	O2	84.46	+	+	+
P. aeruginosa VRFPA04	681.803	1	66.5	O11	86.96	+	+	-
P. aeruginosa	627.851	176	66.1	O6	90.54	+	+	+
P. aeruginosa 148	664.374	128	66.1	O11	90.93	+	+	-
P. aeruginosa ID4365	677.663	172	66.1	O7	91.74	+	+	+
P. aeruginosa C2773C	671.772	200	65.9	O6	93.96	+	+	+


Raw reads
Assembly
Assembly
BLASTn: OSA database
1 OSA match (>95%)
>1 OSA match (>95%)
No OSA match (>95%)
Mono-typeable
Multi-typeable
Non-typeable
Output:
- Serogroup(s)/best hit
- Query OSA fasta file


O11/O17: 14.11%
O12: 11.25%
O13/O14: 0.36%
O2/O16: 1.7%
O10/O19: 2.86%
O1: 7.5%
NT: 2.32%
O9: 1.25%
O7/O8: 1.16%
O3: 25.54%
O6: 21.7%
O4: 4.46%
O5/O18/O20: 5.8%


O11/O17: 1.51%
O10/O19: 7.18%
O1: 17.96%
O6: 32.89%
NT: 0.19%
O7/O8: 15.31%
O13/O14: 2.65%
O2/O16: 3.59%
O9: 0.38%
O3: 6.81%
O4: 1.89%
O5/O18/O20: 9.64%



14
12
10
8
6
4
2
0
Number of isolates
0-20
20-40
40-60
60-80
80-95
OSA coverage (%)
O1
O6
O7
O2
O11
O13