Unexpected diversity in the mobilome of a *Pseudomonas aeruginosa* strain isolated from a dental unit waterline revealed by SMRT Sequencing

Journal:	<i>Genome</i>
Manuscript ID	gen-2017-0239.R1
Manuscript Type:	Article
Date Submitted by the Author:	22-Jan-2018
Complete List of Authors:	Vincent, Antony; Universite Laval Institut de Biologie Integrative et des Systemes Charette, Steve; Universite Laval Institut de Biologie Integrative et des Systemes, Barbeau, Jean; Universit�� de Montr��al
Is the invited manuscript for consideration in a Special Issue? :	N/A
Keyword:	<i>Pseudomonas aeruginosa</i> , dental unit waterline, mobilome, insertion sequences, SMRT sequencing

SCHOLARONETM
Manuscripts

Unexpected diversity in the mobilome of a *Pseudomonas aeruginosa* strain isolated from a dental unit waterline revealed by SMRT Sequencing

Antony T. Vincent^{1,2,3}, Steve J. Charette^{*1,2,3} and Jean Barbeau⁴

1. Institut de biologie intégrative et des systèmes (IBIS), Université Laval, Quebec City, Canada

2. Centre de recherche de l’Institut universitaire de cardiologie et de pneumologie de Québec (CRIUCPQ), Quebec City, Canada

3. Département de biochimie, de microbiologie et de bio-informatique, Université Laval, Quebec City, Canada

4. Département de stomatologie, Faculté de Médecine Dentaire, Université de Montréal, Montreal City, Canada

Antony T. Vincent : antony.vincent.1@ulaval.ca

Steve J. Charette : steve.charette@bcm.ulaval.ca

Jean Barbeau : jean.barbeau@umontreal.ca

*Steve J. Charette, Institut de Biologie Intégrative et des Systèmes (IBIS), Pavillon Charles-Eugène-Marchand, 1030 avenue de la Médecine, Université Laval, Quebec City, QC, Canada, G1V 0A6. E-mail: Steve.Charette@bcm.ulaval.ca

Abstract

The Gram-negative bacterium *Pseudomonas aeruginosa* is found in several habitats, both natural and human-made, and is particularly known for its recurrent presence as a pathogen in the lungs of patients suffering from cystic fibrosis, a genetic disease. Given its clinical importance, several major studies have investigated the genomic adaptation of *P. aeruginosa* in lungs and its transition as acute infections become chronic. However, our knowledge about the diversity and adaptation of the *P. aeruginosa* genome to non-clinical environments is still fragmentary, in part due to the lack of accurate reference genomes of strains from the numerous environments colonized by the bacterium. Here, we used PacBio long-read technology to sequence the genome of PPF-1, a strain of *P. aeruginosa* isolated from a dental unit waterline. Generating this closed genome was an opportunity to investigate genomic features that are difficult to accurately study in a draft genome (contigs state). It was possible to shed light on putative genomic islands, some shared with other reference genomes, new prophages, and the complete content of insertion sequences. In addition, four different group II introns were also found, including two characterized here and not listed in the specialized group II intron database.

Keywords: dental unit waterlines, *Pseudomonas aeruginosa*, PacBio, mobilome, insertion sequences, introns, genomic islands, prophages

Graphical abstract: Investigation of mobile genetic elements of *Pseudomonas aeruginosa* strain PPF-1

Introduction

The Gram-negative bacterium *Pseudomonas aeruginosa* is known to efficiently colonize several environments, including both clinical settings and natural habitats (Moradali et al. 2017). This pathogen, which is the causative agent of major infections and for which several strains are multi-resistant to antibiotics (Cabot et al. 2016; Gellatly and Hancock 2013), is considered by the World Health Organization as one of the three bacteria with the highest priority for the development of new drugs. *P. aeruginosa* is particularly known for its recurrent presence in the lungs of patients that suffer from cystic fibrosis (CF), a genetic disorder that causes, among other symptoms, an obstruction of airways (Cutting 2015).

Given the crucial importance of *P. aeruginosa* in a medical context, genomes from clinical strains have been thoroughly investigated, permitting the discovery, for example, that *P. aeruginosa* can modify its genome, and the elements that are encoded there, to establish a chronic infection in CF lungs (Winstanley et al. 2016) or even to adapt to non-CF bronchiectasis lungs (Hilliam et al. 2017). However, the method that *P. aeruginosa* uses to adapt to other environments, such as those that are non-clinical, is still poorly characterized.

Dental unit waterlines (DUWLs) are environments known to be colonized by a vast array of bacterial species, including *P. aeruginosa* (Abdouchakour et al. 2015; Barbeau et al. 1996). Recent studies investigated phenotypes (Ouellet et al. 2015) and genomes (Vincent et al. 2017) of DUWL *P. aeruginosa* strains. These studies introduced new insights about how this bacterium adapts to the DUWL environment and revealed several unexpected characteristics. For example, we found altered genes involved in quorum sensing (*lasR*) and in the O-specific antigen (*wzx*), caused by copies of the insertion sequence *ISPa11*. However, all investigated genomic sequences

were in a draft state (in many contigs), and this restricted robust genomic studies to small-non-duplicated mutations.

The inability to conduct investigations of the repeated elements is a serious issue, considering that one of the major characteristics was the presence of *ISPa11*, which has been found in several copies. It is even more a problem when considering that large-scale features such as the presence of genomic islands and prophages cannot be properly investigated in draft genomes (Fadeev et al. 2016; Soares et al. 2016). It is crucial to have a complete accurate genome sequence of a DUWL *P. aeruginosa* strain to have a clearer idea of the genomic features of the strains from this environment that could be missed when investigating genome sequences in a draft state.

With long-read technology SMRT Sequencing from PacBio, we have succeeded in obtaining this crucial piece of the puzzle, and have sequenced the DNA of PPF-1, a *P. aeruginosa* strain isolated from a DUWL. A complete accurate chromosome sequence has been obtained, allowing us to shed light on insertion sequences, prophages and genomic islands found in this genome. Interestingly, autocatalytic introns of group II were also found. The high-quality genome sequence of PPF-1 could be a reference for other studies interested in genomes from non-clinical *P. aeruginosa*.

Materials and methods

P. aeruginosa strain PPF-1 was isolated from a DUWL at the Université de Montréal (Canada) dental clinic (Ouellet et al. 2015). The strain's DNA was extracted by phenol/chloroform using the protocol proposed by Pacific Biosciences (<http://www.pacb.com>). The DNA was then sequenced with the Pacific Biosciences RS II system at the Génome Québec Innovation Centre (McGill University, Montreal, Canada). The sequencing reads were *de novo* assembled using

HGAP (Chin et al. 2013) through SMRT Analysis (protocol RS_HGAP_Assembly.3). The chromosome sequence was circularized using Circlator version 1.5.1 (Hunt et al. 2015). The sequence was polished by mapping the Illumina reads from another study (Vincent et al. 2017) using a combination of BWA version 0.7.12-r1039 (Li and Durbin 2009), SAMtools version 1.3 (Li et al. 2009) and Pilon version 1.22 (Walker et al. 2014).

The sequence of PPF-1 was annotated using the NCBI's *Prokaryotic Genome Annotation Pipeline* (Tatusova et al. 2016). Prophages, insertion sequences, and genomic islands were annotated through PHASTER (Arndt et al. 2016), ISSaga2 (Varani et al. 2011) and IslandViewer 4 (Bertelli et al. 2017), respectively. Genes associated with genomic islands were grouped into functional categories using eggNOG-mapper (Huerta-Cepas et al. 2017). Group II introns were found by performing blastn searches between group II intron sequences from the *Database for bacterial group II introns* (Candales et al. 2012) and the sequence of PPF-1. All sequences from intron-encoded proteins were downloaded and aligned using MUSCLE version 3.8.31 (Edgar 2004). The resulting alignment was filtered by trimAl version 1.4 using the “automated1” parameter (Capella-Gutiérrez et al. 2009). A molecular phylogeny was performed by Bayesian inference by running five independent chains under the heterogeneous model CAT+GTR for 5,000 cycles with PhyloBayes version 4.1 (Lartillot et al. 2009). A consensus topology was calculated from the saved trees using bpcomp included in the package PhyloBayes after a burn-in of 1000 trees (20%). The largest discrepancy across all bipartitions (maxdiff) was 0.028, meaning that the convergence between the chains was achieved.

The annotated and curated sequence of PPF-1 has now been deposited in GenBank under the accession number CP023316. The genome sequence and its annotation were visualized using Circleator version 1.0.0 (Crabtree et al. 2014). The comparisons between the genome of PPF-1

with the ones of the reference strains PA7 (NC_009656.1), PA14 (NC_008463.1), PAO1 (NC_002516.2) and LESB58 (NC_011770.1) were made using MegaBLAST with default parameters (Morgulis et al. 2008) and visualized using Circleator version 1.0.0.

Results and discussion

General features

The complete genome of *P. aeruginosa* PPF-1, a strain isolated from a DUWL and with unusual phenotypes (Ouellet et al. 2015), was already sequenced by Illumina MiSeq and analyzed in a previous study (Vincent et al. 2017). At the time of that study, the genome of PPF-1 was in a draft state (83 contigs, N50 = 251,118 bp). As indicated in the introduction, it is not easy to investigate the genome architecture and large-duplicated elements of draft sequences. The genome of PPF-1 was consequently sequenced again, using the SMRT technology of PacBio, to obtain a single chromosomal sequence and thus shed light on the features that were left unexplored in the previous study. The general characteristics of the complete PPF-1 genome are shown in Table 1. The differences between the draft genome and the one newly sequenced by PacBio were checked. The tool QUAST (Gurevich et al. 2013) revealed that 99% (6,879,898 bp over 6,930,893 bp) of the closed genome was covered by the draft sequences. Also, a total of 301 mismatches and 14 InDels were discovered. Based on these results, we believe that the vast majority of genomic features, with the exceptions of large-repeated elements, were properly analyzed previously (Vincent et al. 2017).

Insertion sequences

Insertion sequences are self-transposable elements widespread in bacterial genomes (Siguier et al. 2014). Since insertion sequences are usually repeated, they are known to be a major cause of assembly breakages (Ricker et al. 2012; Vincent et al. 2014), making them difficult to study in draft genomes (Tanaka et al. 2017).

It has already been inferred that the genome of the PPF-1 strain, isolated from a DUWL, harbored several copies of insertion sequence *ISPa11* (Vincent et al. 2017). Copies of this insertion sequence interrupted several genes. These interrupted genes include *lasR*, which encodes a master regulator of the *P. aeruginosa* quorum sensing (Papenfort and Bassler 2016) and *wzx*, which produces a putative O-antigen flippase (Liu et al. 1996).

Knowing that insertion sequences were already known to have altered genetic features of PPF-1, its closed genome was an opportunity for a more robust investigation of the complete repertoire of insertion sequences of this genome. In addition to the 12 *ISPa11*s, 8 other complete insertion sequences and 3 partial were also found (Table 2 and Table S1). These insertion sequences are distributed in 8 types, in 5 families.

It is important to note that there is some confusion concerning the name *ISPa11*. The *ISPa11* found in PPF-1, also listed elsewhere (Dean and Goldberg 2000), is not the same *ISPa11* as the one listed in the reference database ISfinder (Siguier et al. 2006). The *ISPa11* in ISfinder is among the *IS110* family while the one of PPF-1 is putatively from the *IS30* family due to similarities with other insertion sequences of this family (based on blastn analysis against the ISfinder database). The study that has described this insertion sequence also reported it to be from the *IS30* family (Dean and Goldberg 2000). This example demonstrates the importance of centralized resources to formalize nomenclature and avoid such coincident names.

Except for genes inactivated by *ISPa11* copies, few other genes appear to be clearly inactivated by insertion sequences (Table S1). A copy of IS222 clearly inserted into a known gene, which is coding for a fimbrial protein (GenBank: KYO86255.1), while *ISPa16*, *ISPa32* and *ISPPu1* are adjacent to truncated integrase genes. In these last cases, it is unknown if the insertion sequences played a role in truncation of the integrase genes.

In addition to inactivating genes, insertion sequences are known to promote genome reshaping, and may even influence the bacteria's dependence on its host (Siguier et al. 2014). For example, a study reported that insertion sequences of *P. aeruginosa*, mainly those of the IS3 family, are involved in genome rearrangements (Al-Nayyef et al. 2015). In the genome of PPF-1, two types of ISs, IS222 and *ISPa32*, are from the IS3 family. However, there is no clear evidence that these insertion sequences are implied in large-scale rearrangements in the genome of PPF-1, since four of the five IS222s are in genomic islands, and there is only a single *ISPa32*.

Genomic islands

Having the complete accurate genome of PPF-1 was an opportunity to investigate large-scale features, such as genomic islands (GEIs), which are difficult to study in draft genomes (Fadeev et al. 2016; Soares et al. 2016). Numerous putative GEIs were found in the genome of PPF-1 (Figure 1). A total of 1074 genes were predicted to be associated with GEIs by the tool IslandViewer 4 (Bertelli et al. 2017) (Supplementary file S1). Since some GEIs are overlapping, there are 733 non-redundant genes encoded by GEIs. Of these 733 sequences, 562 were further grouped into functional categories (Figure S1). The most represented categories are S (Function unknown), L (Replication, recombination, and repair), M (Cell wall/membrane/envelope biogenesis) and K (Transcription). It is also interesting to note that several genes that code for

176 drug exporters, in addition to genes putatively involved in resistance to copper, arsenic, and
177 mercury, have been found in GEIs and may play a role in defense and adaptation (Supplementary
178 file S1). In addition, several GEIs encode restriction-modification systems that could have helped
179 them to bypass bacterial host defenses (Murphy et al. 2013).

180 When comparing the genome sequence of PPF-1 with the ones of the *P. aeruginosa* reference
181 strains PAO1, PA14, LESB58 and PA7, several GEIs have been found to be unique to PPF-1.
182 The sequence of PA7 shares the most GEIs with PPF-1. PA14 shares the second-most GEIs with
183 PPF-1 (Figure 1). Strains of *P. aeruginosa* are known to be distributed in three large phylogenetic
184 groups (Freschi et al. 2015; Stewart et al. 2014; Vincent et al. 2017), PAO1 and LESB58 being in
185 group 1, PA14 in group 2, and PA7 in group 3. The PPF-1 strain, as with other isolates from
186 DUWLs, are in group 2 along with PA14 (Vincent et al. 2017). It was consequently expected that
187 both PPF-1 and PA14 would share similar genomic features. However, members of group 3, such
188 as PA7, are evolutionarily far from those of groups 1 and 2 and were even sometimes qualified as
189 outliers (Roy et al. 2010). This suggests that at least some GEIs have the potential to be
190 transferred between strains from different groups of *P. aeruginosa*, even with those from the
191 divergent group 3.

192 Interestingly, several of the regions that we predicted to be GEIs in the genome of PPF-1 are also
193 predicted by IslandViewer 4 to be GEIs in the genomes of PA7 and PA14, however, without
194 clear homology (Figure 2). It is known that genomes of *P. aeruginosa*, as many other bacteria,
195 have hotspots for the integration of horizontally acquired genetic elements (Mathee et al. 2008;
196 Oliveira et al. 2017). The present result suggests that the regions where the GEIs inserted into the
197 PPF-1 genome could be hotspots for insertion of such mobile elements.

Prophages

Knowing the high number of GEIs, it was interesting to find out if prophages, which are the DNA of phages that are integrated into the bacterial chromosome, were present in the genome of PPF-1. These phage elements are known to have altered the lifestyle and the genome of several bacterial species (Brüssow et al. 2004). The approaches include, but are not limited to, making initially avirulent strains virulent, through lysogenic conversion (Fortier and Sekulovic 2013). Prophages are also genomic elements that could be difficult to investigate in draft genomes (Kingsford et al. 2010). We predicted that two prophages, one complete and one questionable, were present in the genome of PPF-1 (Figure 1). The questionable prophage (18,502 bp) was found to be almost identical (97–99% of identity over 96–100% of query cover) in PAO1, PA14 and LESB58, while more distant (83% of identity over 84% of query cover) in PA7. The fact that the homologous region is more distant in PA7 was expected, since this strain is known to be highly divergent at the nucleotide level when compared to other *P. aeruginosa* strains (Freschi et al. 2015; Vincent et al. 2017). This observation also suggests that the integration of this putative prophage could have occurred a long time ago. However, when compared specifically against the viral sequences database, the best hit corresponded to a *Pseudomonas* phage of the *Myoviridae* family, phi CTX (GenBank Y13918.1), with an identity of 76% over only 8% of the genome.

The sequence of the complete prophage, compared to the one of the questionable prophage, is much more distant from sequences that can be found in GenBank. The best hits were against the genome of *P. aeruginosa* strains DN1 (CP017099.1), H5708 (CP008859.2) and USDA-ARS-USMARC-41639 (CP013989.1), where the prophage sequence was found at 97–98% of identity over 63–68% of query cover. Fragments of this prophage were also found in the sequence of the reference strains PAO1, PA14 and LESB58, but in a lesser proportion (97–98% of identity over

221 17–19% of query cover). Interestingly, the genome of PA7 harbors a region with a higher identity
222 match (96% of identity over 65% of query cover). Although it is impossible to draw robust
223 conclusions based on the temporality of acquisition of these phage regions, the fact that the
224 genome of a phylogenetically distant strain such as PA7 possesses a putative prophage similar to
225 the one found in the genome of PPF-1 suggests recent integration events. When compared to the
226 virus sequences in GenBank, this putative complete prophage shares 98% and 92% of identity for
227 20% of query length with the *Pseudomonas* phages YMC11/07/P54_PAE_BP (KU310943.1) and
228 YMC11/02/R656 (KT968831.1), respectively. Both YMC11/07/P54_PAE_BP and
229 YMC11/02/R656 are from South Korea and of the Siphoviridae family.

230 Finally, it is interesting to note that this putative complete prophage sequence includes some
231 genomic island regions that we predicted (Figure 1). It is unclear if genomic islands inserted into
232 this prophage, since this was a region with low conservative pressure, given that the prophage is
233 also a part of the accessory genome, or if this is the result of a limitation of the prediction
234 method.

235 Group II introns

236 An overview of the annotated features made it possible to find several genes that code for group
237 II intron reverse transcriptase/maturase. Group II introns are autocatalytic and consist of a
238 ribozyme RNA and, usually, an open reading frame [intron-encoded protein (IEP)], which
239 encodes a reverse transcriptase/maturase (Pyle 2016). These mobile elements were listed in
240 genomes of bacteria, archaeobacteria, mitochondria, and chloroplasts (Zimmerly and Semper
241 2015). In 2011, group II introns were estimated to be present in around 25% of the sequenced
242 bacterial genomes (Lambowitz and Zimmerly 2011). According to the group II intron database

(Candales et al. 2012), group II introns are classified based on their RNA types (IIA, IIB, and IIC) and on the phylogenetic clustering of their IEPs [which could be bacterial (A, B, C, D, E, F and G), mitochondrial-like (ML), or chloroplast-like (CL1 and CL2)].

The PPF-1 genome harbors two group II introns (P.ae.I2 and P.ae.I3) present in the database mentioned above, which have already been listed in some genomes of *P. aeruginosa* (Figure 3). Two new introns that do not appear in the database were also found in the genome of PPF-1 (named here as P.ae.I4 and P.ae.I5) by sequence homology with those present in the database and then by manual curation. A molecular phylogeny of 317 IEP sequences allowed us to determine that P.ae.I4 and P.ae.I5 are from the class CL1 and have type IIB1 RNA (Figure S2). Interestingly, P.ae.I5 is present twice in the genome of PPF-1. Blastn analyses against the nr/nt database of the NCBI revealed that the introns found in the genome of PPF-1 are also present in other genomes of *P. aeruginosa* (Figure 3 and Table S2). However, no closed genome from GenBank have all the introns found in the genome of PPF-1.

According to the statistics kept at IMG (the Integrated Microbial Genome Database) (Markowitz et al. 2012), around 88% of the bacterial genomes that are known so far are still in a draft state. Although draft genomes are valuable for several applications, it is complicated to analyze their architectures and mobile genetic elements (such as insertion sequences, genomic islands and prophage regions) (Fadeev et al. 2016; Kingsford et al. 2010; Ricker et al. 2012). Fortunately, there is a democratization of what is called third-generation-sequencing technologies. This new technological advance makes it much more possible for researchers to complete the bacterial genomes they are working on (Bleidorn 2016) and to have a clearer idea of the features encoded by these genomes (Li et al. 2016).

By generating, completing, and polishing the genome of *P. aeruginosa* PPF-1, a strain isolated from a DUWL, we were able to shed light on a complex mobilome that comprises insertion sequences, genomic islands, prophages, and group II introns.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery grant RGPIN-2014-04595 to S.J.C]. A.T.V holds an Alexander Graham Bell Canada Graduate Scholarships from the NSERC. SJC is a research scholar from the Fonds de Recherche du Québec en Santé.

References

- Abdouchakour F., Dupont C., Grau D., Aujoulat F., Mournetas P., Marchandin H., et al. 2015. *Pseudomonas aeruginosa* and *Achromobacter* sp. clonal selection leads to successive waves of contamination of water in dental care units. *Appl. Environ. Microbiol.* **81**(21): 7509–7524. doi:10.1128/AEM.01279-15.
- Al-Nayyef H., Guyeux C., Petitjean M., Hocquet D., Bahi J.M. 2015. Relation between insertion sequences and genome rearrangements in *Pseudomonas aeruginosa*. In: *Bioinformatics and Biomedical Engineering* (F. Ortuño and I. Rojas, eds). Springer International Publishing:Cham. 426–437.
- Arndt D., Grant J.R., Marcu A., Sajed T., Pon A., Liang Y., et al. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**(W1): W16–W21. doi:10.1093/nar/gkw387.
- Barbeau J., Tanguay R., Faucher E., Avezard C., Trudel L., Côté L., et al. 1996. Multiparametric

- 286 analysis of waterline contamination in dental units. Appl. Environ. Microbiol. **62**(11): 3954–
287 3959.
- 288 Bertelli C., Laird M.R., Williams K.P., Lau B.Y., Hoad G., Winsor G.L., et al. 2017.
289 IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. Nucleic
290 Acids Res. **45**(W1): W30–W35. doi:10.1093/nar/gkx343.
- 291 Bleidorn C. 2016. Third generation sequencing: technology and its potential impact on
292 evolutionary biodiversity research. Syst. Biodivers. **14**(1): 1–8.
293 doi:10.1080/14772000.2015.1099575.
- 294 Brüssow H., Canchaya C., Hardt W-D. 2004. Phages and the evolution of bacterial pathogens:
295 from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. **68**(3):
296 560–602. doi:10.1128/MMBR.68.3.560-602.2004.
- 297 Cabot G., Zamorano L., Moyà B., Juan C., Navas A., Blázquez J., et al. 2016. Evolution of
298 *Pseudomonas aeruginosa* antimicrobial resistance and fitness under low and high mutation
299 supply rates. Antimicrob. Agents Chemother. **60**(3): 1767–1778. doi:10.1128/AAC.02676-
300 15.
- 301 Candales M.A., Duong A., Hood K.S., Li T., Neufeld R.A.E., Sun R., et al. 2012. Database for
302 bacterial group II introns. Nucleic Acids Res. **40**: D187–D190. doi:10.1093/nar/gkr1043.
- 303 Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: A tool for automated
304 alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**(15): 1972–
305 1973. doi:10.1093/bioinformatics/btp348.
- 306 Chin C-S., Alexander D.H., Marks P., Klammer A.A., Drake J., Heiner C., et al. 2013.

- 307 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.
308 Nat. Methods **10**(6): 563–569. doi:10.1038/nmeth.2474.
- 309 Crabtree J., Agrawal S., Mahurkar A., Myers G.S., Rasko D.A., White O. 2014. Circleator:
310 Flexible circular visualization of genome-associated data with BioPerl and SVG.
311 Bioinformatics **30**(21): 3125–3127. doi:10.1093/bioinformatics/btu505.
- 312 Cutting G.R. 2015. Cystic fibrosis genetics: from molecular understanding to clinical application.
313 Nat. Rev. Genet. **16**(1): 45–56. doi:10.1038/nrg3849.
- 314 Dean C.R., Goldberg J.B. 2000. The *wbpM* gene in *Pseudomonas aeruginosa* serogroup O17
315 resides on a cryptic copy of the serogroup O11 O antigen gene locus. FEMS Microbiol. Lett.
316 **187**(1): 59–63. doi:10.1016/S0378-1097(00)00175-0.
- 317 Edgar R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high
318 throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340.
- 319 Fadeev E., De Pascale F., Vezzi A., Hübner S., Aharonovich D., Sher D. 2016. Why close a
320 bacterial genome? The plasmid of *Alteromonas macleodii* HOT1A3 is a vector for inter-
321 specific transfer of a flexible genomic island. Front. Microbiol. **7**: 248. doi
322 10.3389/fmicb.2016.00248.
- 323 Fortier L-C., Sekulovic O. 2013. Importance of prophages to evolution and virulence of bacterial
324 pathogens. Virulence **4**(5): 354–365. doi:10.4161/viru.24498.
- 325 Freschi L., Jeukens J., Kukavica-Ibrulj I., Boyle B., Dupont M.J., Laroche J., et al. 2015. Clinical
326 utilization of genomics data produced by the international *Pseudomonas aeruginosa*
327 consortium. Front. Microbiol. **6**: 1036. doi:10.3389/fmicb.2015.01036.

- 328 Gellatly S.L., Hancock R.E.W. 2013. *Pseudomonas aeruginosa*: New insights into pathogenesis
329 and host defenses. *Pathog. Dis.* **67**(3): 159–173. doi:10.1111/2049-632X.12033.
- 330 Gurevich A., Saveliev V., Vyahhi N., Tesler G. 2013. QUAST: quality assessment tool for
331 genome assemblies. *Bioinformatics* **29**(8): 1072–1075. doi:10.1093/bioinformatics/btt086.
- 332 Hilliam Y., Moore M.P., Lamont I.L., Bilton D., Haworth C.S., Foweraker J., et al. 2017.
333 *Pseudomonas aeruginosa* adaptation and diversification in the non-cystic fibrosis
334 bronchiectasis lung. *Eur. Respir. J.* **49**(4): 1602108. doi:10.1183/13993003.02108-2016.
- 335 Huerta-Cepas J., Forslund K., Coelho L.P., Szklarczyk D., Jensen L.J., Von Mering C., et al.
336 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-
337 mapper. *Mol. Biol. Evol.* **34**(8): 2115–2122. doi:10.1093/molbev/msx148.
- 338 Hunt M., De Silva N., Otto T.D., Parkhill J., Keane J.A., Harris S.R. 2015. Circlator: automated
339 circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**: 294.
340 doi:10.1186/s13059-015-0849-0.
- 341 Kingsford C., Schatz M.C., Pop M. 2010. Assembly complexity of prokaryotic genomes using
342 short reads. *BMC Bioinformatics* **11**: 21. doi:10.1186/1471-2105-11-21.
- 343 Lambowitz A.M., Zimmerly S. 2011. Group II introns: Mobile ribozymes that invade DNA. *Cold*
344 *Spring Harb. Perspect. Biol.* **3**(8): 1–19. doi:10.1101/cshperspect.a003616.
- 345 Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: A Bayesian software package for
346 phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**(17): 2286–2288.
347 doi:10.1093/bioinformatics/btp368.

- 348 Li G., Shen M., Le S., Tan Y., Li M., Zhao X., et al. 2016. Genomic analyses of multidrug
349 resistant *Pseudomonas aeruginosa* PA1 resequenced by single-molecule real-time
350 sequencing. Biosci. Rep. **36**(6): e00418. doi:10.1042/BSR20160282.
- 351 Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
352 Bioinformatics **25**(14): 1754–1760. doi:10.1093/bioinformatics/btp324.
- 353 Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., et al. 2009. The Sequence
354 Alignment/Map format and SAMtools. Bioinformatics **25**(16): 2078–2079.
355 doi:10.1093/bioinformatics/btp352.
- 356 Liu D., Cole R.A., Reeves P.R. 1996. An O-antigen processing function for Wzx (RfbX): A
357 promising candidate for O-unit flippase. J. Bacteriol. **178**(7): 2102–2107.
- 358 Markowitz V.M., Chen I-M.A., Palaniappan K., Chu K., Szeto E., Grechkin Y., et al. 2012. IMG:
359 the integrated microbial genomes database and comparative analysis system. Nucleic Acids
360 Res. **40**: D115–D122. doi: 10.1093/nar/gkr1044
- 361 Mathee K., Narasimhan G., Valdes C., Qiu X., Matewish J.M., Koehrsen M., et al. 2008.
362 Dynamics of *Pseudomonas aeruginosa* genome evolution. Proc. Natl. Acad. Sci. **105**(8):
363 3100–3105. doi:10.1073/pnas.0711982105.
- 364 Moradali M.F., Ghods S., Rehm B.H.A. 2017. *Pseudomonas aeruginosa* lifestyle: A paradigm
365 for adaptation, survival, and persistence. Front. Cell. Infect. Microbiol. **7**: 39.
366 doi:10.3389/fcimb.2017.00039.
- 367 Morgulis A., Coulouris G., Raytselis Y., Madden T.L., Agarwala R., Schäffer A.A. 2008.
368 Database indexing for production MegaBLAST searches. Bioinformatics **24**(16): 1757–

- 369 1764. doi:10.1093/bioinformatics/btn322.
- 370 Murphy J., Mahony J., Ainsworth S., Nauta A., van Sinderen D. 2013. Bacteriophage orphan
371 DNA methyltransferases: Insights from their bacterial origin, function, and occurrence.
372 Appl. Environ. Microbiol. **79**(24): 7547–7555. doi:10.1128/AEM.02229-13.
- 373 Oliveira P.H., Touchon M., Cury J., Rocha E.P.C. 2017. The chromosomal organization of
374 horizontal gene transfer in bacteria. Nat. Commun. **8**(1): 841. doi:10.1038/s41467-017-
375 00808-w.
- 376 Ouellet M.M., Leduc A., Nadeau C., Barbeau J., Charette S.J. 2015. *Pseudomonas aeruginosa*
377 isolates from dental unit waterlines can be divided in two distinct groups, including one
378 displaying phenotypes similar to isolates from cystic fibrosis patients. Front. Microbiol. **5**:
379 802. doi: 10.3389/fmicb.2014.00802.
- 380 Papenfort K., Bassler B.L. 2016. Quorum sensing signal-response systems in Gram-negative
381 bacteria. Nat. Rev. Microbiol. **14**(9): 576–88. doi:10.1038/nrmicro.2016.89.
- 382 Pyle A.M. 2016. Group II intron self-splicing. Annu. Rev. Biophys. **45**: 183–205.
383 doi:10.1146/annurev-biophys-062215-011149.
- 384 Ricker N., Qian H., Fulthorpe R.R. 2012. The limitations of draft assemblies for understanding
385 prokaryotic adaptation and evolution. Genomics **100**(3): 167–175.
386 doi:10.1016/j.ygeno.2012.06.009.
- 387 Roy P.H., Tetu S.G., Larouche A., Elbourne L., Tremblay S., Ren Q., et al. 2010. Complete
388 genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7.
389 PLoS One **5**(1): e8842. doi:10.1371/journal.pone.0008842.

- 390 Siguier P., Gourbeyre E., Chandler M. 2014. Bacterial insertion sequences: Their genomic impact
391 and diversity. *FEMS Microbiol. Rev.* **38**(5): 865–891. doi:10.1111/1574-6976.12067.
- 392 Siguier P., Perochon J., Lestrade L., Mahillon J., Chandler M. 2006. ISfinder: the reference
393 centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**: D32–D36.
394 doi:10.1093/nar/gkj014.
- 395 de Castro Soares S., de Castro Oliveira L., Jaiswal A.K., Azevedo V. 2016. Genomic Islands: an
396 overview of current software tools and future improvements. *J. Integr. Bioinform.* **13**(1):
397 301. doi:10.2390/biecoll-jib-2016-301.
- 398 Stewart L., Ford A., Sangal V., Jeukens J., Boyle B., Caim S., et al. 2014. Draft genomes of
399 twelve host adapted and environmental isolates of *Pseudomonas aeruginosa* and their
400 position in the core genome phylogeny. *Pathog. Dis.* **71**(1): 20–25. doi:10.1111/2049-
401 632X.12107.
- 402 Sullivan M.J., Petty N.K., Beatson S.A. 2011. Easyfig: a genome comparison visualiser.
403 *Bioinformatics* **27**(7): 1009–1010. doi:10.1093/bioinformatics/btr039.
- 404 Tanaka K.H., Vincent A.T., Emond-Rheault J-G., Adamczuk M., Frenette M., Charette S.J. 2017.
405 Plasmid composition in *Aeromonas salmonicida* subsp. *salmonicida* 01-B526 unravels
406 unsuspected type three secretion system loss patterns. *BMC Genomics* **18**(1): 528.
407 doi:10.1186/s12864-017-3921-1.
- 408 Tatusova T., Dicuccio M., Badretdin A., Chetvernin V., Nawrocki E.P., Zaslavsky L., et al. 2016.
409 NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**(14): 6614–6624.
410 doi:10.1093/nar/gkw569.

- 411 Varani A.M., Siguier P., Gournayre E., Charneau V., Chandler M. 2011. ISSaga is an ensemble of
412 web-based methods for high throughput identification and semi-automatic annotation of
413 insertion sequences in prokaryotic genomes. *Genome Biol.* **12**(3): R30. doi:10.1186/gb-
414 2011-12-3-r30.
- 415 Vincent A.T., Boyle B., Derome N., Charette S.J. 2014. Improvement in the DNA sequencing of
416 genomes bearing long repeated elements. *J. Microbiol. Methods* **107**: 186–188.
417 doi:10.1016/j.mimet.2014.10.016.
- 418 Vincent A.T., Freschi L., Jeukens J., Kukavica-Ibrulj I., Emond-Rheault J-G., Leduc A., et al.
419 2017. Genomic characterisation of environmental *Pseudomonas aeruginosa* isolated from
420 dental unit waterlines revealed the insertion sequence *ISPa11* as a chaotropic element.
421 *FEMS Microbiol. Ecol.* **93**(9): fix106. doi: 10.1093/femsec/fix106
- 422 Walker B.J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., et al. 2014. Pilon: An
423 integrated tool for comprehensive microbial variant detection and genome assembly
424 improvement. *PLoS One* **9**(11): e112963. doi:10.1371/journal.pone.0112963.
- 425 Winstanley C., O'Brien S., Brockhurst M.A. 2016. *Pseudomonas aeruginosa* evolutionary
426 adaptation and diversification in cystic fibrosis chronic lung infections. *Trends Microbiol.*
427 **24**(5): 327–337. doi:10.1016/j.tim.2016.01.008.
- 428 Zimmerly S., Semper C. 2015. Evolution of group II introns. *Mob. DNA* **6**: 7.
429 doi:10.1186/s13100-015-0037-5.
- 430
- 431

432 **Table 1.** General features of the genome of PPF-1

Length (bp)	6,930,893
GC (%)	65.91
CDSs	6501
rRNA genes	12
tRNA genes	62
Introns	5
Insertion sequences (complete + partial)	20 + 3
Density (gene per kb)	0.918

433
434

Draft

435 **Table 2.** ISs found in the PPF-1 genome

IS	Family	Complete	Partial	Host
ISPa11	IS30 ^a	12	0	<i>P. aeruginosa</i>
IS222	IS3	3	2	<i>P. aeruginosa</i>
ISPa16	IS5	1	0	<i>P. aeruginosa</i>
ISPa32	IS3	1	0	<i>P. aeruginosa</i>
ISPa37	IS30	1	0	<i>P. aeruginosa</i>
ISPpu1	IS630	1	0	<i>P. putida</i>
ISPst12	IS5	1	0	<i>P. stutzeri</i>
ISPst3	IS21	0	1	<i>P. stutzeri</i>
Total		20	3	

436 a. Inferred by sequence homology against ISfinder database (Siguier et al. 2006)

437

Figure captions

Figure 1. Map of the chromosome of *P. aeruginosa* PPF-1. The outer red ring represents the GC percent. The two blue inner rings represent the genes encoded on the forward and the reverse strands, respectively. The green rectangles represent the predicted genomic islands, while the two orange-framed rectangles represent the prophages. Black arrows represent the position of group II introns. The pink, dark blue, yellow and purple rings are the homologous regions between the genome of the strain PPF-1 and the ones of reference strains PA7 (NC_009656.1), PA14 (NC_008463.1), PAO1 (NC_002516.2) and LESB58 (NC_011770.1), respectively. Finally, the teal ring represents the GC skew.

Figure 2. Multiple alignments of the genome sequences of PA7 (NC_009656.1), PPF-1 (CP023316) and PA14 (NC_008463.1). Direct and inverted homologous regions are represented by the orange and blue zones, respectively, while GEIs are represented by the green rectangles. The figure was obtained using EasyFig version 2.2.2 (Sullivan et al. 2011).

Figure 3. Content in group II introns found in the genome of PPF-1 and their absence/presence in complete genomes available in GenBank. Gray, light blue and dark blue rectangles represent introns that are absent, present, and present twice in complete genomes of *P. aeruginosa* available in GenBank. The class and RNA type are indicated for all introns. Introns marked with a “*” (P.ae.I4 and P.ae.I5) are not in the group II intron database (Candales et al. 2012) and were found by the present study.

Captions for Supplementary figures and tables (gen-2017-0239.R1Suppla)

Table S1. Detailed information on the insertion sequences found in the genome of PPF-1

Table S2. Identity of group II introns found in the genome of PPF-1 with those available in GenBank (identified with MegaBLAST)

Figure S1. Distribution of 562 genes associated to genomic islands into functional categories.

Figure S2. Molecular phylogeny of 317 intron-encoded proteins (IEPs) as described in the main text.

Draft

Intron