Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing

L A Snyder[1,2], N J Loman[1], L A Faraj[3], K Levi[4], G Weinstock[5], T C Boswell[4], M J Pallen (m.pallen@warwick.ac.uk)[6], D A Ala'Aldeen[3]

1. Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, Birmingham, United Kingdom
2. Kingston University, Kingston upon Thames, Surrey, United Kingdom (current affiliation)
3. Molecular Bacteriology and Immunology Group, School of Life Sciences, University of Nottingham, Nottingham, United Kingdom
4. Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom
5. Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, United States
6. Division of Microbiology and Infection, Warwick Medical School, University of Warwick, Coventry, United Kingdom

Although previous bacterial typing methods have been informative about potential relatedness of isolates collected during outbreaks, next-generation sequencing has emerged as a powerful tool to not only look at similarity between isolates, but also put differences into biological context. In this study, we have investigated the whole genome sequence of five *Pseudomonas aeruginosa* isolates collected during a persistent six-year outbreak at Nottingham University Hospitals National Health Service (NHS) Trust – City Campus, United Kingdom. Sequencing, using both Roche 454 and Illumina, reveals that most of these isolates are closely related. Some regions of difference are noted between this cluster of isolates and previously published genome sequences. These include regions containing prophages and prophage remnants such as the serotype-converting bacteriophage D3 and the cytotoxin-converting phage phi CTX. Additionally, single nucleotide polymorphisms (SNPs) between the genomic sequence data reveal key single base differences that have accumulated during the course of this outbreak, giving insight into the evolution of the outbreak strain. Differentiating SNPs were found within a wide variety of genes, including *lasR*, *nrdG*, *tadZ*, and *algB*. These have been generated at a rate estimated to be one SNP every four to five months. In conclusion, we demonstrate that the single base resolution of whole genome sequencing is a powerful tool in analysis of outbreak isolates that can not only show strain similarity, but also evolution over time and potential adaptation through gene sequence changes.

Introduction

Pseudomonas aeruginosa is a Gram-negative bacterium that is widespread in the environment and engages in a wide variety of interactions with eukaryotic host organisms. It is a common opportunistic pathogen in humans, causing a broad range of infections in community and healthcare settings. The most serious manifestations of infection include bacteraemia (particularly in neutropenic patients), pneumonia (particularly in cystic fibrosis patients and critically ill patients), urinary tract infections and wound infections (especially in patients with burn injuries) [1,2]. This organism is intrinsically resistant to many antibiotics and in recent years resistance has emerged to what were previously antimicrobial agents of choice [3,4].

As *P. aeruginosa* is ubiquitous in the environment and also forms part of the resident microbiota of many patients, molecular typing methods are required to identify the sources and routes of transmission within the healthcare setting. Current molecular epidemiological typing systems, such as pulsed-field gel electrophoresis (PFGE) and variable-number tandem repeat (VNTR) analyses, have provided evidence of clonal outbreaks of *P. aeruginosa* affecting clinics or hospitals across the United Kingdom (UK) [5,6]. For example, such approaches have identified several clones (the Liverpool, Midlands 1 and Manchester strains) affecting multiple cystic fibrosis patients in the UK [5-7]. However, these typing methods are of limited resolution when compared to whole genome sequencing. In addition, traditional typing methods fail to provide any insights into the biology of outbreak strains, or show how they might be evolving over time.

A prolonged *P. aeruginosa* outbreak at Nottingham University Hospitals National Health Service (NHS) Trust - City Campus (UH-NHST-CC) that lasted from

2001 to 2007 illustrates some of these issues [8]. Although the outbreak has been brought under control, one to two cases are still seen each year, so it has not been completely eliminated. Here, we have seen a pattern of apparently endemic infection, associated with a single clonal lineage, which was isolated from over 40 patients and also from the hospital environment, as evidenced in previous studies of the isolates [8] and ongoing active surveillance and molecular typing.

The outbreak was first detected when infections caused by five multi-resistant isolates were noted on an 18-bed bone marrow transplant unit at UH-NHST-CC in April 2001. These isolates were resistant to ceftazidime (CTZ), ciprofloxacin (CIP), gentamicin (GEN), piperacillin-tazobactam (TZP), and the carbapenems, but susceptible to amikacin (AMK) and colistin (COL). Surveillance was enhanced, laboratory records were reviewed to identify other cases, and environmental sampling was undertaken. Over the next two years, further isolates from the critical care unit and a new bone marrow transplant unit at the same hospital were identified, initially on the basis of the same antibiogram. As a result of the infection control programmes initiated in response to these outbreaks, infections were identified in 32 haematology patients [8]. Some patients acquired the multi-resistant *P. aeruginosa* without temporal overlap with other positive patients, suggesting an environmental source. Infection control investigations at the time suggested that the outbreak was associated with colonisation of hand washing basins.

Initial molecular epidemiological typing was performed by PFGE using restriction enzyme *Spe*I [9,10] and identified a single outbreak strain among clinical and environmental isolates [8]. Similarity between all isolates was ›85% and these formed a single group in cluster analysis, defined as the Nottingham/Trent cluster. In addition, the recommendations of Tenover et al. [10] were followed, which stipulate that isolates belonging to the same strain should differ by a maximum of four fragments. Isolates were additionally typed by random amplification of polymorphic DNA (RAPD) [11] and all demonstrated the same pattern [8]. It was found that in critical care, tracheostomy was the strongest risk factor for acquisition and that the tracheostomy inner tubes were being cleaned in hand wash basins that were found to be contaminated. The correct fitting of waste trap heaters, the removal of one persistently colonised sink, enhanced surveillance, and a review of equipment and sink cleaning, particularly in the case of tracheostomy equipment helped to bring the outbreak under control. In clinical haematology, oral colistin was introduced for selective decontamination of the gut during neutropenia and use of ciprofloxacin, to which the strain is resistant, was reduced for neutropenic prophylaxis.

Whole genome sequencing, facilitated by the advent of high-throughput approaches, brings the promise

of single-base-pair resolution between isolates, making it the ultimate molecular typing method for bacteria. Several recent studies have shown that analysis of single nucleotide polymorphisms (SNPs) in bacterial genomes provides a means of determining relatedness between epidemiologically linked isolates and tracking bacterial evolution over periods of months to years [12-16]. In particular, several studies have tracked the genetic changes that emerge during long-term colonisation of cystic fibrosis patients [17,18]. In addition, genome sequencing provides a strain-specific list of protein-coding genes, which, through comparisons with the genes sets of other strains from the same species can provide insights into the unique biology of an epidemic strain. We therefore applied high-throughput sequencing to five isolates from the UH-NHST-CC hospital outbreak to determine whether genome sequencing can inform our understanding of the epidemiology and biology of a hospital pathogen in this setting.

Methods

Microbiology

Isolates of *P. aeruginosa* were collected from the microbiological specimens of patients cared for at the Queen's University Medical Centre over a six-year period from January, 2002 to December, 2007. Isolates were frozen on beads and stored at -80°C. *P. aeruginosa* isolates were sent to the Health Protection England's (HPE) Laboratory of HealthCare Associated Infection in London for analysis by PFGE. Four isolates from different time points in the outbreak were selected for genome sequencing in this study on the basis of PFGE and RAPD results, as well as an additional isolate from June 2001 (Table 1). This additional isolate was identified as part of the look-back exercise initiated after the first outbreak; its PFGE profile was established by the HPE. The antimicrobial susceptibilities of isolates were tested using the disk diffusion method against AMK, aztreonam (AZT), CTZ, COL, CIP, GEN, meropenem (MER), tobramycin (TOB) and TZP and according to

British Society for Antimicrobial Chemotherapy (BSAC) guidelines [19].

Isolation of genomic DNA and high-throughput sequencing

Genomic DNA was obtained from colony-purified *P. aeruginosa*, using the DNeasy DNA extraction kit according to the manufacturer's instructions (Qiagen). DNA from isolate PANOT106 was sequenced using whole genome shotgun and 2.5-kilobase paired-end 454 protocols in accordance with the manufacturer's instructions (Roche) at the University of Liverpool's Centre for Genomic Research. Isolates PANOT106, PANOTK11, PANOT101, PANOT340, and PANOT738 were sequenced one sample per lane using the Illumina GA2, running 35 cycles with paired-end libraries (~250 base-pair insert size). Sequencing on the Illumina GA2 platform was carried out at the sequencing centre of the University of Washington at St. Louis.

Genome assembly

Roche 454 reads for isolate PANOT106 were assembled de *novo* with Newbler v2.3 (Roche) to create a reference assembly, using the default settings with –g enabled to produce contig graphs for use later in the variant detection pipeline. Nesoni 0.51 (Victorian Bioinformatics Consortium) was used on the Illumina reads for PANOT106 to correct errors in the reference assembly, particularly short indels in regions with low depth of coverage or in homopolymeric tracts. Nesoni parameters were: depth 10; purity 0.8; ambiguity-codes 1; fidelity 1.0; monogamous 1; trim 0; max-pair-sep 250; only-pairs 1; suffix1 /1; suffix2 /2; same-dir 0; circular 0; savehits 0. Bowtie v0.12.0 [20] was used to align the PANOT106 Illumina reads to complete *P. aeruginosa* genomes available in GenBank, using parameters --solexa -quals -S -m 1 -p 4. The progressive Mauve component of Mauve (version 2.3.1) was used to map the PANOT106 reference assembly to the genome sequence of *P. aeruginosa* PAO1.

Illumina sequences for isolates PANOTK11, PANOT101, PANOT340, and PANOT738 were assembled using SOAPdenovo 1.03 (http://soap.genomics.org.cn/soapdenovo.html). Assemblies were individually annotated using the xBASE annotation pipeline [21-26].

Creation of a well-validated single nucleotide polymorphism (SNP) set

A series of filters were applied to the sequence data to create a set of high-confidence SNP variants. Bowtie v0.12.0 was used to align the Illumina data for each of the five sequenced strains to the PANOT106 reference assembly and variants extracted using samtools with parameters –cv –N *n* to create set ‹VARIANTS›. The value of *n* was set to the number of isolates in the sample.

Putative variants were filtered with samtools.pl varFilter to create set ‹FILTER1›. varFilter was run with the following parameters: -Q 25 -d 10 -D 100 -G 25 -w 10 -W 10 -N 2 -l 30, equating to the following filtering options: Minimum root mean square mapping quality for SNPs, 25; Minimum root mean square mapping quality for gaps, 10; Minimum read depth, 10; Maximum read depth, 90; Minimum indel score for nearby SNP filtering, 25; SNP within INT bp around a gap to be filtered, 10; Window size for filtering dense SNPs, 10; Max number of SNPs in a window, 2; Window size for filtering adjacent gaps, 30.

These variant calls were further filtered to a set with SNP quality of 30 (phred-scale) to create set ‹FILTER2›. The position of each variant call was related to the *P. aeruginosa* strain PAO1 genome sequence by reference to the Mauve alignment. Variant calls with no corresponding alignment in strain PAO1 were discarded, creating set ‹FILTER3›. Variant calls which occurred in contigs in the reference assembly which were not included in the scaffold assembly and thus were likely to be repetitive were further excluded from analysis, resulting in ‹FILTER4›. Variant calls which occurred in contigs which had coverage depth greater or less than two standard deviations from the mean were excluded, resulting in ‹FILTER5›. Finally, variant calls with a variant frequency of less than 0.9 were discarded. Variant frequency was determined by dividing the alignment depth by the number of reads supporting the variant call.

Analysis of gene content

Protein-coding sequences (CDSs) in the PANOT106 assembly were identified using Gene Locator and Interpolated Markov ModelER (GLIMMER) [23]. Protein Basic Local Alignment Search Tool (BLASTP) [21] was used to compare the set of CDSs in PANOT106 with the predicted protein repertoires of other completed *P. aeruginosa* genome sequences. Regions of difference (RODs) between the PANOT106 and the known *P. aeruginosa* pan-genome (as defined by complete genomes in GenBank) were identified on the basis of the following criteria: the ROD must contain ≥ 10 strain-specific CDSs, which have no orthologues in other genomes (as defined by mutual best-hit analysis), with the ROD interrupted by no more than five genes already present in the pan-genome. Orthologues were computed using OrthoMCL v2.0 with default parameters [27]. Singletons (coding sequences which do not have orthologues in other outbreak isolates) were verified by translated nucleotide databases BLAST (TBLASTN) searches against the original sequence. Additional information with regards to the annotated features made use of the *Pseudomonas* Genome Database [28].

Results

Assembly of the genome sequence data

Using the Roche 454 next-generation genome sequencing platform, 6,720,954 bases were generated from the DNA extracted from isolate PANOT106. Reads were assembled, using Newbler, into 1,059 contigs, which were further assembled into 19 scaffolds using

Isolate	Total reads	Reads per pair	Bases	Coverage	Scaffolds	Maximum scaffold size	Bases in scaffolds	Average scaffold size	N50 scaffold size	N90 scaffold size
PANOTK11	11,936,530	5,968,265	417,778,550	~70X	393	168,419	6,857,254	11,024	41,132	9,556
PANOT101	15,007,136	7,503,568	525,249,760	~88X	189	288,005	6,823,034	17,494	110,879	25,874
PANOT106	15,392,542	7,696,271	538,738,970	~90X	202	414,841	6,823,529	18,196	81,409	20,375
PANOT340	10,684,280	5,342,140	373,949,800	~62X	756	161,292	6,821,477	6,363	20,384	3,728
PANOT738	14,983,686	7,491,843	524,429,010	~87X	294	225,595	6,818,630	14,001	51,282	12,127

NHS: National Health Service.

paired-end sequence data to determine the relative positions and distances of contigs. The N50 of the scaffolds is 825,048 and the average paired end distance is 2,365 bases, with a standard deviation of 591 bases.

The next-generation sequencing data generated on the Illumina platform for isolates PANOT106, PANOTK11, PANOT101, PANOT340, and PANOT738, is summarised in Table 2. Coverage with the 35 base Illumina reads ranged from 62X to 90X. The data for isolate PANOT106 was used to correct the assembly of this genome sequence based on the 454 data. This resulted in 283 total edits to the original assembled sequence.

Antibiogram differences

The only detected difference in antibiotic resistance amongst the five isolates was AMK resistance in PANOTK11 (data not shown). Resistance to aminoglycosides is typically associated with aminoglycoside-modifying enzymes. Each of the outbreak strains carry identical aminoglycoside 3'-phosphotransferase type IIb coding sequences homologous to the *Pseudomonas* strain PAO1 CDS with accession PA4119. Additionally there were no SNPs detected upstream to suggest a change in promoter in PANOTK11. The alternative resistance-conferring aminoglycoside 4'-O-adenylyltransferase IIb [29] was not found in any of the outbreak strains. We therefore conclude that AMK resistance in PANOTK11 is probably due to impermeability or efflux, although even if true, this fails to explain why resistance to other aminoglycosides was unaffected [30].

Novel sequences within the genomes of the outbreak isolates

There are six large regions of difference between PANOT106 and the *P. aeruginosa* pan-genome (Figure panel A-F), five of which appear to be prophages or prophage remnants. In particular, PANOT106 contains a region that resembles the serotype-converting *P. aeruginosa* bacteriophage D3 and a region that encodes homologues of phage structural proteins from phi CTX,

a cytotoxin-converting phage of *P. aeruginosa*. The only non-prophage region of difference contains copper- and arsenic-resistance genes and is homologous to a similar region in the genome of *Stenotrophomonas* sp. SKA14.

PANOTK11 has a 48 kb sequence that is not present in the other strains sequenced here (Figure panel G). This region was assembled into a single contig and was annotated to contain 24 CDSs. Although this region has no homologue in any of the other strains sequenced here, it is homologous to sequence data from isolate PACS171b [31], which was isolated from the oropharynx of a nine month-old cystic fibrosis patient [32]. Of the 48 kb, all but 13 kb share similarity with isolate PACS171b; these 35 kb mostly contain hypothetical genes and a putative restriction-modification system and putative sigma-54 transcriptional regulator. Within the 13 kb region that is different from isolate PACS171b, isolate PANOTK11 contains a member of the DEAD box helicase family, an OLD family endonuclease, and a hypothetical gene.

No novel sequences were found within the genomes of PANOT101, PANOT340, and PANOT738, in comparison with the *P. aeruginosa* pan-genome.

Flagellum operon

All five of the outbreak strains are missing a portion of the flagellum operon compared to that present in *P. aeruginosa* strains PAO1 and LESB58. The corresponding CDS PA1088, encoding flagellar glycosyl transferase FgtA is not present in the outbreak isolates sequenced here. The *fliC* (PA1092) gene and CDSs PA1092–PA1096 are present, although these are divergent in their nucleotide sequences.

Genome-wide single nucleotide polymorphism (SNP) analysis of outbreak strains

The PANOT106 sequence, assembled from Illumina and 454 sequence data, aligns well with the *P. aeruginosa* strains PAO1 (95.3% sequence identity),

Regions of sequence difference between Nottingham *Pseudomonas aeruginosa* isolates and the *P. aeruginosa* pan-genome

Regions of sequence difference between isolate PANOT106 and the *P. aeruginosa* pan-genome

Region of sequence difference between isolate PANOTK11 and the *P. aeruginosa* pan-genome
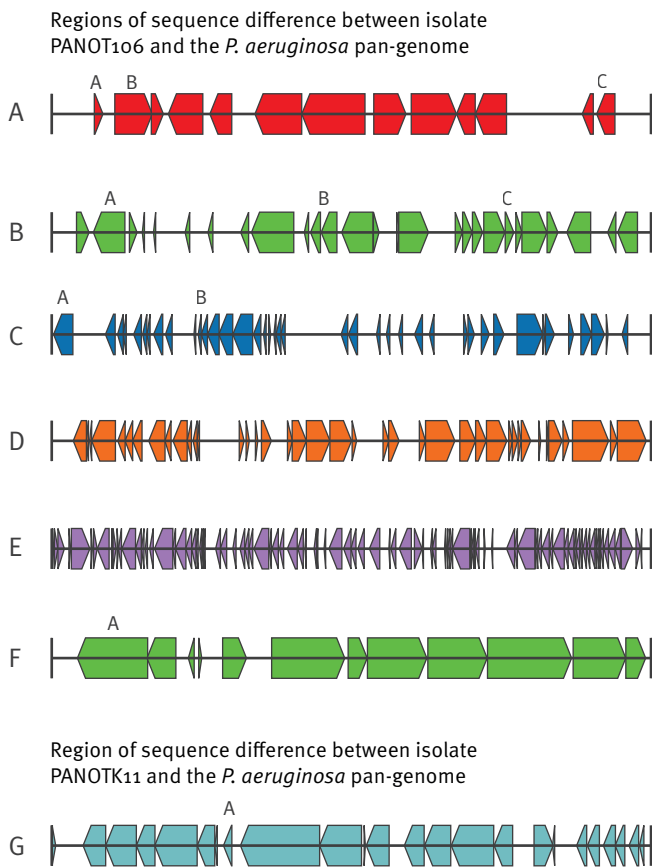
Arrowed blocks show annotated coding regions, indicating their direction of transcription.

Panel A. A 17.2 kb prophage region containing a chromate resistance protein (A), chromate transporter (B), and phage structural proteins (C).

Panel B. A 29.4 kb copper/arsenate resistance region containing a heavy metal Resistance-Nodulation-cell Division (RND) efflux outer membrane protein (A), a copper resistance operon (B), and arsenate resistance genes (C).

Panel C. A 32.1 kb lambda-like prophage region containing a phage integrase (A) and alginate biosynthesis regulatory protein (B).

Panel D. A 35.3 kb prophage region.

Panel E. A Phi-CTX-like prophage region.

Panel F. A 21.4 kb serine protease containing region of difference containing subtilisin-related serine protease (A).

Panel G. A 47.9 kb region of difference from PANOTK11 containing a tellurium resistance protein (A).

LESB58 (92.5%), and PA14 (92.4%) genome sequences and less well with PA7 (77.8%). Comparisons of the Ilumina sequences from isolates PANOTK11, PANOT101, PANOT340, and PANOT738 with the PANOT106 assembly generated a large number of variant calls. Several hundred variants were discarded during the filtering of variants to identify high-quality informative SNPs that vary between the outbreak isolates and which are associated with 37 polymorphic loci (Table 3). Twenty-four of the SNPs result in non-synonymous changes to the associated encoded amino acid in the PAO1-annotated coding region (Table 4). This high frequency of non-synonymous changes (24 of 37) suggests that some of the SNPs represent adaptive changes.

Nineteen of the SNPs separate PANOTK11 from the other four UH-NHST-CC Nottingham isolates (Table 3). None of these nineteen SNPs are found in any of the other four outbreak isolates, which all share greater similarity to the reference strain PAO1 at these SNP loci. These SNPs in this isolate, the 48 kb region present only in this isolate, and the AMK resistance of this isolate indicate that PANOTK11, despite being identified as related to the other outbreak isolates as part of the Nottingham/Trent cluster via PFGE and RAPD, is not part of the same outbreak and cannot represent the index case. Despite the earlier typing results, PANOTK11 is a phylogenetic outlier when compared to other outbreak isolates on the genomic level. Interestingly, one of these PANOTK11-specific SNPs (PNSNP12) maps to *lasR* (Table 4), encoding a transcriptional regulator that responds to a homoserine lactone signal to activate expression of *P. aeruginosa* virulence factors through quorum sensing. Mutations in *LasR* have been reported in several contexts, including adaptation to the airways of cystic fibrosis patients [33]. PNSNP12 does not appear to be within the helix-turn-helix DNA binding domain of the encoded protein according to the available sequence and functional information in UniProt [35], and to predictions by the helixturnhelix (HTH) software tool [34] but SNP mutations in other *LasR* regions, including the ligand binding domain, have been reported previously [36]. Another of the PANOTK11 SNPs (PNSNP34) maps to *algB* (Table 4), encoding an NtrC-like response regulator, which controls alginate biosynthesis in mucoid *P. aeruginosa* [37]. In this case the SNP falls within the helix-turn-helix motif predicted by HTH [34] and described previously [38]. This SNP may therefore impact upon the virulence of PANOTK11 and differences in alginate biosynthesis should be explored.

PANOT101 and PANOT106 were isolated a year apart (2002 and 2003, respectively) and share many of the same non-synonymous SNPs, apart from five identified here (Table 3). PNSNP10, within a sensor/response regulator hybrid (PA0928), and PNSNP25, within *nrdG*, in PANOT106 are also shared with isolates PANOT340 and PANOT738 (Tables 3 and 4). PNSNP11 and PNSNP13 in a hypothetic protein (PA0938) and two-component sensor (PA1458), respectively, are only found in PANOT106 in this study. Likewise, PNSNP26 in PANOT101 (Table 3) is only found within this isolate in this study. PNSNP26 is within *speC* encoding ornithine decarboxylase (PA4519; Table 4) part of the biosynthetic pathway to produce putrescine, required for growth, but loss of function of this protein can be compensated through an alternative putrescine biosynthesis pathway using *speA* [39].

Comparison of well-validated single nucleotide polymorphisms in the *Pseudomonas aeruginosa* Nottingham outbreak isolates

SNP ID	*P. aeruginosa* Nottingham outbreak isolates				
	PANOT101	PANOT106	PANOT340	PANOT738	PANOTK11
PNSNP1	G	G	G	G	A
PNSNP2	T	T	G	G	T
PNSNP3	C	C	C	A	C
PNSNP4	T	T	T	T	C
PNSNP5	G	G	T	T	G
PNSNP6	A	A	A	A	G
PNSNP7	A	A	A	A	T
PNSNP8	T	T	T	T	C
PNSNP9	A	A	A	A	G
PNSNP10	T	G	G	G	T
PNSNP11	A	T	A	A	A
PNSNP12	G	G	G	G	A
PNSNP13	C	T	C	C	C
PNSNP14	G	G	G	G	C
PNSNP15	T	T	T	T	A
PNSNP16	G	G	G	T	G
PNSNP17	C	C	C	C	T
PNSNP18	C	C	C	A	C
PNSNP19	C	C	T	T	C
PNSNP20	A	A	A	A	C
PNSNP21	C	C	C	C	G
PNSNP22	C	C	C	A	C
PNSNP23	A	A	G	G	A
PNSNP24	A	A	A	A	C
PNSNP25	T	G	G	G	T
PNSNP26	A	C	C	C	C
PNSNP27	G	G	G	T	G
PNSNP28	A	A	A	A	G
PNSNP29	T	T	T	T	C
PNSNP30	C	C	C	C	T
PNSNP31	G	G	G	A	G
PNSNP32	A	A	C	C	A
PNSNP33	A	A	A	C	A
PNSNP34	G	G	G	G	C
PNSNP35	C	C	C	A	C
PNSNP36	G	G	G	G	T
PNSNP37	T	T	T	T	G

ID: identity; SNP: single nucleotide polymorphism.

The two subsequent strains, PANOT340 and PANOT738, isolated in 2005 and 2007, respectively, show a progressive increase in the number of SNPs relative to PANOT106. These three isolates share two non-synonymous SNPs (Table 3), one, PNSNP10, within a sensor/response regulator hybrid (PA0928, *gacS*) and

one, PNSNP25, in a class III (anaerobic) ribonucleoside-triphosphate reductase activating protein gene (PA1919, *nrdG*). Five shared non-synonymous SNPs (PNSNP2, PNSNP5, PNSNP19, PNSNP23, and PNSNP32) are present in isolates PANOT340 and PANOT738, but not in the other outbreak isolates. In isolate PANOT738, an additional eight SNPs are present, three leading to non-synonymous changes (Table 3) and five in intergenic regions. This progressive accumulation of SNPs is consistent with an ancestor-descendant relationship. The acquisition of five SNPs in two and a third years (5 SNPs in 28 months, or 5.6 months per SNP) and a further eight SNPs in two and two third years (8 SNPs in 32 months, or 4 months per SNP), with 13 over a span of five years from January 2003 to December 2007 (13 SNPs in 60 months, 4.6 months per SNP) suggests that a new SNP is acquired in this lineage on average every four to five months.

The five non-synonymous SNPs in common between isolates PANOT340 and PANOT738 include PNSNP32 which is associated with a type II citrate synthase (PA1580, *gltA*), PNSNP2 in a hypothetical protein that may be a thioesterase (PA3130), PNSNP5 in a putative methylated-DNA--protein-cysteine methyltransferase (PA3596), PNSNP23 in a probable two-component response regulator (PA5364), and PNSNP19 in an ATP-dependent protease peptidase subunit of heat shock protein HslV (PA5053, *hslV*). The small form of the citrate synthase protein encoded by *gltA* is involved in the Krebs cycle and is expressed in excess during stationary phase growth [40]. Whether PNSNP32 has some bearing on the stationary phase growth rate of isolates PANOT340 and PANOT738 in comparison with isolate PANOT106 requires further investigation, as this may impact upon survival of the bacteria in the hospital setting. PNSNP5 is within a CDS potentially encoding a methylated-DNA--protein-cysteine methyltransferase, which is involved in the repair of alkylated DNA (EC 2.1.1.63). Sensitivity testing for alkylating agents of the outbreak isolates may be warranted in light of the PNSNP5 data.

The remaining eight SNPs present only in isolate PANOT738 are associated with non-synonymous changes in three different CDSs and five intergenic regions. The three SNPs in CDSs are PNSNP33 found within a putative class III pyridoxal phosphate-dependent aminotransferase (PA0530), PNSNP27 in *tadZ* (PA4303), and PNSNP16 in a putative transcriptional regulator (PA2846). The CDS for *tadZ* is within the widespread colonisation island (WCI) encoding tight adherence pili. The genes of the WCI are essential for biofilm formation and colonisation [41], which are important virulence factors in *P. aeruginosa*. The PNSNP27 within *tadZ* may therefore influence the success of isolate PANOT738 in forming biofilms and colonising the host, medical devices, or even medical surfaces such as the persistently colonised sink. Isolate PANOT738 also contains PNSNP16 that is within PA2846, which is one of several predicted transcriptional (PA1430)

SNP ID[a]	Isolate SNP[b]	PAO1[c]	Position in PAO1[d]	PAO1 locus[e]	PAO1 CDS annotation	PAO1 amino acid	SNP amino acid
PNSNP2	(C)	A	3512351	PA3130	Hypothetical protein	V	G
PNSNP4	(G)	A	4128733	PA3687	Phosphoenolpyruvate carboxylase	W	R
PNSNP5	(A)	C	4032021	PA3596	Putative methylated-DNA--protein-cysteine methyltransferase	C	F
PNSNP7	(A)	T	4000131	PA3568	Putative acetyl-coa synthetase	I	F
PNSNP9	G	A	827118	PA0758	Hypothetical protein	I	T
PNSNP10	G	T	1014510	PA0928	Sensor/response regulator hybrid	T	P
PNSNP11	T	A	1028593	PA0938	Hypothetical protein	Q	L
PNSNP12	A	G	1558547	PA1430	Transcriptional regulator LasR	G	D
PNSNP13	T	C	1588334	PA1458	Two-component sensor	R	C
PNSNP14	C	G	1590208	PA1459	Chemotaxis-specific methylesterase	G	R
PNSNP15	A	T	1635647	PA1505	Molybdenum cofactor biosynthesis protein A	L	Q
PNSNP16	(A)	C	3201052	PA2846	Putative transcriptional regulator	L	I
PNSNP17	(A)	G	3007141	PA2656	Putative two-component sensor	L	F
PNSNP19	(A)	G	5692646	PA5053	ATP-dependent protease peptidase subunit	G	D
PNSNP23	(C)	T	6031991	PA5364	Putative two-component response regulator	N	S
PNSNP25	G	T	2093908	PA1919	Ribonucleoside-triphosphate reductase activating protein, NrdG	T	P
PNSNP26	A	C	5060237	PA4519	Ornithine decarboxylase	G	C
PNSNP27	T	G	4827417	PA4303	TadZ	P	Q
PNSNP29	T	C	5173073	PA4615	Putative oxidoreductase	E	K
PNSNP32	(G)	T	1719347	PA1580	Type II citrate synthase	D	A
PNSNP33	(G)	T	589009	PA0530	Putative class III pyridoxal phosphate-dependent aminotransferase	L	R
PNSNP34	(G)	C	6174644	PA5483	Two-component response regulator AlgB	L	V
PNSNP36	T	G	2686727	PA2402	Peptide synthase	S	Stop
PNSNP37	G	T	2686790	PA2402	Peptide synthase	D	A

CDS: protein coding sequence; ID: identity; SNP: single nucleotide polymorphism; Stop: termination codon.

[a] SNP ID from Table 3.
[b] SNP from one of the isolates as listed in Table 3. Reverse complementary bases, in parenthesis, have been entered where the PAO1 strand is negative relative to the data in Table 3.
[c] Base at this SNP location within the PAO1 genome sequence.
[d] Position of the SNP within the PAO1 genome sequence.
[e] CDS locus identifier from the PAO1 genome sequence.

and response regulators (PA0928, PA1458, PA2656, PA5364, PA5483) that contain SNPs in these outbreak isolates, suggesting that differences in regulation may play a role in the phenotypes observed.

Discussion

Here, we have shown that genome sequence data obtained with high-throughput sequencing technologies can provide novel insights into the epidemiology and biology of hospital isolates of *P. aeruginosa*. One isolate, PANOTK11, was determined to be an outlier and not part of the same outbreak lineage as the other four

isolates. This is based on genomic level sequence differences; previous typing with PFGE and RAPD placed all five of these isolates into a single lineage.

Amongst the related isolates, we have demonstrated that *P. aeruginosa* lineages from a hospital setting contain a sufficient number of SNPs to allow the construction of an epidemiological narrative — an important benchmarking exercise that lays the foundations for future investigations into the fine-grained genomic epidemiology of this troublesome pathogen.

In addition, genome sequencing has highlighted the existence of strain-specific gene clusters that might underpin the tenacious survival of these strains within our hospital over many years. Several SNPs which may confer phenotypic differences between the isolates have been identified and bear further investigation, particularly as some may play a role in survival of isolates in the hospital. These include SNPs within PA1580 that could influence stationary phase growth rate and impact on the survival of the bacteria in the hospital setting, within PA3596 that could influence the sensitivity to alkylating agents, and within PA4303 that may be important for biofilm formation.

This study illustrates the potential of the first generation of high-throughput sequencing platforms in the investigation of hospital outbreaks. Relentless improvements in cost and performance (particularly those delivered recently by bench top sequencing [42]) as well as data analysis software, are taking these technologies ever closer to the clinical microbiology laboratories. With a price tag now comparable to the cost of traditional molecular typing, bacterial genome sequencing is now poised to impact on clinical practice.

Acknowledgements

Authors' contributions

DAA, MJP, NJL, GW conceived of and designed the study; DAA, LAS, KL, TCB carried out clinical, epidemiological and microbiological analyses; GW supervised genome sequencing; NJL, LAF and LAS performed bioinformatics analyses; MJP, DAA, NJL, LAS wrote the paper.

References

1. Fujitani S, Sun HY, Yu VL, Weingarten JA. Pneumonia due to Pseudomonas aeruginosa: part I: epidemiology, clinical diagnosis, and source. Chest. 2011;139(4):909-19.
http://dx.doi.org/10.1378/chest.10-0166
PMid:21467058

2. Grisaru-Soen G, Lerner-Geva L, Keller N, Berger H, Passwell JH, Barzilai A. Pseudomonas aeruginosa bacteremia in children: analysis of trends in prevalence, antibiotic resistance and prognostic factors. Pediatric Infect Dis J. 2000;19(10):959-63.
http://dx.doi.org/10.1097/00006454-200010000-00003

3. Hirsch EB, Tam VH. Impact of multidrug-resistant Pseudomonas aeruginosa infection on patient outcomes. Expert Rev Pharmacoecon Outcomes Res. 2010;10(4):441-51.
http://dx.doi.org/10.1586/erp.10.49
PMid:20715920 PMCid:PMC3071543

4. Lister PD, Wolter DJ, Hanson ND. Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. Clin Microbiol Rev. 2009;22(4):582-610.
http://dx.doi.org/10.1128/CMR.00040-09
PMid:19822890 PMCid:PMC2772362

5. 5.Smart CH, Scott FW, Wright EA, Walshaw MJ, Hart CA, Pitt TL, et al. Development of a diagnostic test for the Midlands 1 cystic fibrosis epidemic strain of Pseudomonas aeruginosa. J Med Microbiol. 2006;55(Pt 8):1085-91.
http://dx.doi.org/10.1099/jmm.0.46604-0
PMid:16849729

6. Turton JF, Turton SE, Yearwood L, Yarde S, Kaufmann ME, Pitt TL. Evaluation of a nine-locus variable-number tandem-repeat scheme for typing of Pseudomonas aeruginosa. Clin Microbiol Infect. 2010;16(8):1111-6. Epub 2009/09/08.

7. Scott FW, Pitt TL. Identification and characterization of transmissible Pseudomonas aeruginosa strains in cystic fibrosis patients in England and Wales. J Med Microbiol. 2004;53(Pt 7):609-15.
http://dx.doi.org/10.1099/jmm.0.45620-0
PMid:15184530

8. Gee TM. Epidemiological Investigation and Control of a Hospital Wide Outbreak of Multi-drug Resistant Pseudomonas aeruginosa [Thesis submitted to Royal College of Pathologist (United Kingdom) as partial fulfillment of MRCPath Part II]. Work undertaken at Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom; 2004.

9. Kaufmann M. Pulsed Field Gel Electrophoresis. In Woodford N Johnson AP, editors. Methods in Molecular Medicine, Vol. 15: Molecular Bacteriology: Protocols and Clinical Applications. New York: Humana Press. 1998;33-50.

10. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, et al. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol. 1995;33(9):2233-9.
PMid:7494007 PMCid:PMC228385

11. Vogel L, Jones G, Triep S, Koek A, Dijkshoorn L. RAPD typing of Klebsiella pneumoniae, Klebsiella oxytoca, Serratia marcescens and Pseudomonas aeruginosa isolates using standardized reagents. Clin Microbiol Infect. 1999;5(5):270-6.
http://dx.doi.org/10.1111/j.1469-0691.1999.tb00140.x
PMid:11856266

12. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of Escherichia coli O157:H7 outbreaks. Proc Natl Acad Sci U S A. 2011;108(50):20142-7.
http://dx.doi.org/10.1073/pnas.1107176108
PMid:22135463 PMCid:PMC3250189

13. Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM, Mortiboy D, et al. High-throughput whole-genome sequencing to dissect the epidemiology of Acinetobacter baumannii isolates from a hospital outbreak. J Hosp Infect. 2010;75(1):37-41.
http://dx.doi.org/10.1016/j.jhin.2010.01.012
PMid:20299126

14. Octavia S, Lan R. Single-nucleotide-polymorphism typing and genetic relationships of Salmonella enterica serovar Typhi isolates. J Clin Microbiol. 2007;45(11):3795-801.
http://dx.doi.org/10.1128/JCM.00720-07
PMid:17728466 PMCid:PMC2168493

15. Phetsuksiri B, Srisungngam S, Rudeeaneksin J, Bunchoo S, Lukebua A, Wongtrungkapun R, et al. SNP genotypes of Mycobacterium leprae isolates in Thailand and their combination with rpoT and TTC genotyping for analysis of leprosy distribution and transmission. Jpn J Infect Dis. 2012;65(1):52-6.
 PMid:22274158

16. Schürch AC, Kremer K, Hendriks AC, Freyee B, McEvoy CR, van Crevel R, et al. SNP/RD typing of Mycobacterium tuberculosis Beijing strains reveals local and worldwide disseminated clonal complexes. PLoS one. 2011;6(12):e28365.
http://dx.doi.org/10.1371/journal.pone.0028365
PMid:22162765 PMCid:PMC3230589

17. Bragonzi A, Paroni M, Nonis A, Cramer N, Montanari S, Rejman J, et al. Pseudomonas aeruginosa microevolution during cystic fibrosis lung infection establishes clones with adapted virulence. Am J Respir Crit Care Med. 2009;180(2):138-45
http://dx.doi.org/10.1164/rccm.200812-1943OC
PMid:19423715

18. Jelsbak L, Johansen HK, Frost AL, Thogersen R, Thomsen LE, Ciofu O, et al. Molecular epidemiology and dynamics of Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients. Infect Immun. 2007;75(5):2214-24.
http://dx.doi.org/10.1128/IAI.01282-06
PMid:17261614 PMCid:PMC1865789

19. Andrews JM; BSAC Working Party On Susceptibility Testing ft. BSAC standardized disc susceptibility testing method. J Antimicrob Chemother. 2001;48 Suppl 1:43-57.
http://dx.doi.org/10.1093/jac/48.suppl_1.43
PMid:11420336

20. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
http://dx.doi.org/10.1186/gb-2009-10-3-r25
PMid:19261174 PMCid:PMC2690996

21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.
http://dx.doi.org/10.1093/nar/25.17.3389
PMid:9254694 PMCid:PMC146917

22. Chaudhuri RR, Loman NJ, Snyder LA, Bailey CM, Stekel DJ, Pallen MJ. xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res. 2008;36(Database issue):D543-6.
PMid:17984072 PMCid:PMC2238843

23. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23(6):673-9.
http://dx.doi.org/10.1093/bioinformatics/btm009
PMid:17237039 PMCid:PMC2387122

24. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
http://dx.doi.org/10.1186/gb-2004-5-2-r12
PMid:14759262 PMCid:PMC395750

25. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100-8.
http://dx.doi.org/10.1093/nar/gkm160
PMid:17452365 PMCid:PMC1888812

26. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955-64.
PMid:9023104 PMCid:PMC146525

27. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178-89.
http://dx.doi.org/10.1101/gr.1224503
PMid:12952885 PMCid:PMC403725

28. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Yu NY, et al. Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. Nucleic Acids Res. 2011;39(Database issue):D596-600.
http://dx.doi.org/10.1093/nar/gkq869
PMid:20929876 PMCid:PMC3013766

29. Sabtcheva S, Galimand M, Gerbaud G, Courvalin P, Lambert T. Aminoglycoside resistance gene ant(4')-IIb of Pseudomonas aeruginosa BM4492, a clinical isolate from Bulgaria. Antimicrob Agents Chemother. 2003;47(5):1584-8
http://dx.doi.org/10.1128/AAC.47.5.1584-1588.2003
PMid:12709326 PMCid:PMC153341

30. Poole K. Aminoglycoside resistance in Pseudomonas aeruginosa. Antimicrob Agents Chemother. 2005;49(2):479-87.
http://dx.doi.org/10.1128/AAC.49.2.479-487.2005
PMid:15673721 PMCid:PMC547279

31. Hayden HS, Gillett W, Saenphimmachak C, Lim R, Zhou Y, Jacobs MA, et al. Large-insert genome analysis technology detects structural variation in Pseudomonas aeruginosa clinical strains from cystic fibrosis patients. Genomics. 2008;91(6):530-7.
http://dx.doi.org/10.1016/j.ygeno.2008.02.005
PMid:18445516 PMCid:PMC2587363

32. Ernst RK, D'Argenio DA, Ichikawa JK, Bangera MG, Selgrade S, Burns JL, et al. Genome mosaicism is conserved but not unique in Pseudomonas aeruginosa isolates from the airways of young children with cystic fibrosis. Environ Microbiol. 2003;5(12):1341-9.
http://dx.doi.org/10.1111/j.1462-2920.2003.00518.x
PMid:14641578

33. Hogardt M, Heesemann J. Microevolution of Pseudomonas aeruginosa to a Chronic Pathogen of the Cystic Fibrosis Lung. Curr Top Microbiol Immunology. 2013;358:91-118.http://dx.doi.org/10.1007/82_2011_199
PMid:22311171

34. Dodd IB, Egan JB. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res. 1990;18(17):5019-26.
http://dx.doi.org/10.1093/nar/18.17.5019

35. UniProt Consortium . Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2012;40(Database issue):D71-5.
PMid:22102590 PMCid:PMC3245120

36. D'Argenio DA, Wu M, Hoffman LR, Kulasekara HD, Deziel E, Smith EE, et al. Growth phenotypes of Pseudomonas aeruginosa lasR mutants adapted to the airways of cystic fibrosis patients. Molecular Microbiol. 2007;64(2):512-33.
http://dx.doi.org/10.1111/j.1365-2958.2007.05678.x
PMid:17493132 PMCid:PMC2742308

37. Ma S, Selvaraj U, Ohman DE, Quarless R, Hassett DJ, Wozniak DJ. Phosphorylation-independent activity of the response regulators AlgB and AlgR in promoting alginate biosynthesis in mucoid Pseudomonas aeruginosa. J Bacteriol. 1998;180(4):956-68.
PMid:9473053 PMCid:PMC106978

38. Wozniak DJ, Ohman DE. Pseudomonas aeruginosa AlgB, a two-component response regulator of the NtrC family, is required for algD transcription. J Bacteriol. 1991;173(4):1406-13.
PMid:1899859 PMCid:PMC207277

39. Nakada Y, Itoh Y. Identification of the putrescine biosynthetic genes in Pseudomonas aeruginosa and characterization of agmatine deiminase and N-carbamoylputrescine amidohydrolase of the arginine decarboxylase pathway. Microbiology. 2003;149(Pt 3):707-14.
http://dx.doi.org/10.1099/mic.0.26009-0
PMid:12634339

40. Solomon M, Weitzman PDJ. Occurrence of two distinct citrate synthases in a mutant of Pseudomonas aeruginosa and their growth-dependent variation. FEBS Lett. 1983;155(1):157-60.
http://dx.doi.org/10.1016/0014-5793 (83)80230-0

41. Tomich M, Planet PJ, Figurski DH. The tad locus: postcards from the widespread colonization island. Nat Rev Microbiol. 2007;5(5):363-75.
http://dx.doi.org/10.1038/nrmicro1636
PMid:17435791

42. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30(5):434-9.
http://dx.doi.org/10.1038/nbt.2198
PMid:22522955