Molecular epidemiology of *Pseudomonas aeruginosa* in an unsegregated bronchiectasis cohort sharing hospital facilities with a cystic fibrosis cohort

Philip J. Mitchelmore[1,2], Joanna Randall[3], Matthew J. Bull[4], Karen A. Moore[5], Paul A. O'Neill[5], Konrad Paszkiewicz[5], Eshwar Mahenthiralingam[4], Chris J. Scotton[1], Christopher D. Sheldon[2], Nicholas J. Withers[2], Alan R. Brown[5].

[1]Institute of Biomedical and Clinical Sciences, University of Exeter Medical School, Exeter, UK.
[2]Department of Respiratory Medicine, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK
[3]Department of Microbiology, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK
[4]Organisms and Environment Research Division, Cardiff School of Biosciences, Cardiff University, Cardiff, UK.
[5]Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK.

ONLINE SUPPORTING INFORMATION

METHODS

Recruitment and analysis of NCFB and CF cohorts

All isolates were collected prospectively. The NCFB and CF patients were opportunistically recruited via the submission of samples as part of their routine management. The use of chronic anti-pseudomonal therapy was permitted for inclusion. Informed consent was obtained and samples were sent to the local Microbiology Laboratory for processing according to standard procedures. In brief, sputum samples were treated with an equal volume of Mucolyse (Pro-Lab Diagnostics). Neat digested sputum was used to inoculate blood agar, chocolate agar and *Burkholderia cepacia* agar plates for CF patients, and Sabouraud agar plates for both NCFB and CF patients. Subsequently, 5ml of Maximum Recovery Diluent was added to the digested sputum and inoculated onto blood agar, chocolate Agar, cysteine lactose electrolyte deficient Agar, and mannitol salt agar for both CF and NCFB samples. An initial identification of PA based on colony morphology and a positive oxidase test was subsequently confirmed by VITEK MS Matrix-assisted Laser Desorption/Ionisation-Time of Flight Mass Spectrometry (MALDI-TOF MS) (Biomerieux). When PA was identified, 10 representative colonies were picked per sputum sample (based on colony morphologies) and stored on Microbank^{TM} microbial storage beads (Pro-Lab Diagnostics). If PA was not cultivated from a sample, a future sample from the same patient could be submitted at a subsequent visit. Demographics of recruited PA-positive CF and NCFB patients are detailed in Table S1.

	NCFB	**CF**
Recruitment Period (Month/Year-Month/Year)	07/14-06/15	09/15-01/16
Subjects	46	22
Median Age (in years)	69	27.5
IQR for age	65.8-76	24.5-40.3
Male	10 (21.7%)	10 (45.5%)
Time since 1st PA isolate		
- Less than 1 year	12 (26.1%)	0
- 1-5 years	11 (23.9%)	9 (40.9%)
- 5-9 years	11 (23.9%)	4 (18.2%)
- More than 9 years	12 (26.1%)	9 (40.9%)
Co-pathogens		
- *Staphylococcus aureus*	1 (2.2%)	9 (40.9%)
- *Aspergillus fumigatus*	0	1 (4.5%)
- *Exophiala* species	0	2 (9.1%)
- Other	7 (15.2%)	3 (13.6%)
Antibiotic therapy		
- Current azithromycin use	18 (39.1%)	12 (54.5%)
- Current inhaled anti-pseudomonal use	16 (34.8%)	18 (81.8%)
- Neither azithromycin nor inhaled anti-pseudomonal use	17 (37%)	1 (4.5%)

Table S1. Patient demographics for PA-positive NCFB and CF cohorts.

Recruitment and analysis of non-respiratory cohort

Recruitment and consent were not required for the prospectively collected non-respiratory isolates. No patient details were known to the investigators beyond the disease site (e.g. wound) and whether the specimens were from the community or within the hospital. Confirmation of PA identity was performed as described above.

Genotyping pipeline by RAPD and MLST

RAPD analysis of PA from NCFB and CF cohorts was initially performed one patient at a time to facilitate the visualisation of all unique RAPD profiles within the 10 isolates per patient. The criteria for visually defining a unique RAPD profile was (a) the presence/absence of a major band, (b) the presence/absence of two minor bands, or (c) a difference in one band being major/minor and the presence/absence of another band. From each patient, a representative of each unique profile (based on visual inspection) was taken forward for repeat RAPD analysis within a single batch encompassing representatives from all patients. This same selection of representative isolates was subjected to MLST analysis to enable strain identification in a global context. For the repeat RAPD analysis, RAPD profiles were analysed via microfluidic amplicon separation (Agilent 2100 Bioanalyser), with cluster analysis subsequently performed using Gelcompar II software (Applied Maths), with Pearson's Method and 2% optimisation. A typical RAPD-derived dendrogram is shown in Fig. S1, which depicts all representative NCFB isolates. A 90% similarity cut-off was used to identify clusters of isolates. All isolates depicted in Fig. S1 were subjected to MLST analysis, regardless of whether they were deemed to be unique (<90 % similarity to other isolates) or part of a cluster (≥90% similarity with other isolates).

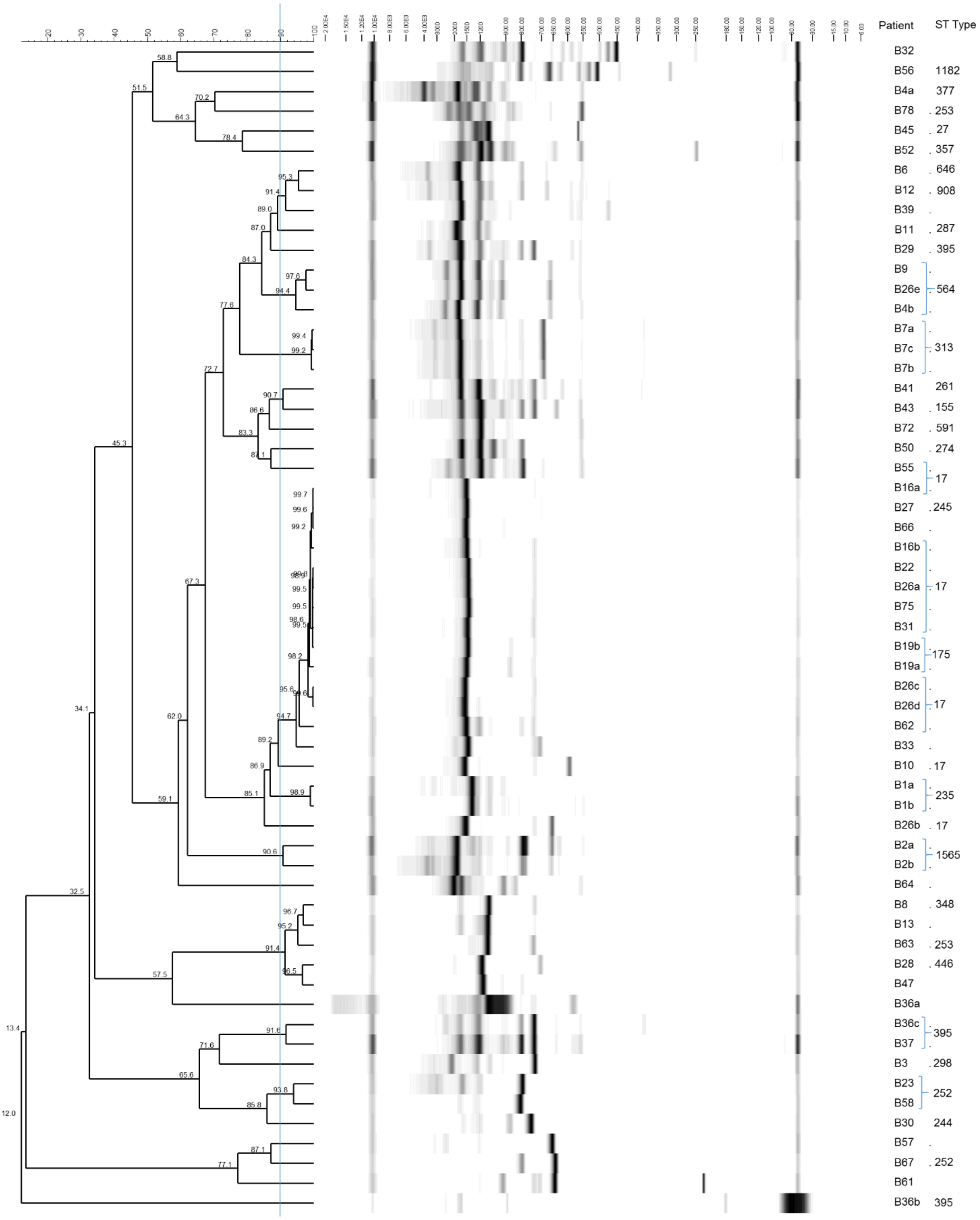

Fig. S1. Dendrogram of RAPD profiles of NCFB PA isolates, derived using microfluidic amplicon separation (Agilent 2100 Bioanalyser). The dendrogram was produced using Gelcompar II (Applied Maths) with Pearson's method and 2% optimisation. A 90% cut-off (denoted by the vertical blue line) was used to identify clusters.

Whole genome sequencing & SNP distance matrices

Data was generated on a single run of an Illumina MiSeq system using 2 x 300bp read lengths. Analysis was carried out using the MRC CLIMB infrastructure [1] using a Virtual Machine pre-installed with the Nullarbor package (https://github.com/tseemann/nullarbor) to generate core genome variants, characterise MLST profiles, resistome, SNP distance matrices and generate a pan-genome report. Nullarbor trims low quality and adaptors present in reads using Trimmomatic [2], Kraken [3] to assign reads to taxonomic groups along with SPAdes [4] to assemble genomes (using the --accurate option). Annotation was performed using Prokka [5] and variants called with Snippy (https://github.com/tseemann/snippy).

Relevant reference genomes obtained from the *Pseudomonas* Genome Database [6] were included in our analysis to enable isolates to be compared to clonally-unrelated representatives of the same sequence type. By cross-reference to the *Pseudomonas aeruginosa* MLST Database (https://pubmlst.org/paeruginosa/; [7]), three genomes were selected for each Sequence Type (ST) of interest (with the exception of ST564 for which no publicly-available genomes were available). Relevant genomes (Table S2) were selected on the basis of them being either complete or with a high Contig N50. In order for these genomes to be incorporated into the Nullarbor package, the genomes were processed using the wgsim package (v 0.3.2) to generate simulated short paired-end reads for re-assembly. The following parameters were used to generate 2x300bp reads with zero changes with respect to the reference genome -e 0.000000 -d 600 -1 300 -2 300 -r 0.000000 –R 0.0000 -X 0.0000 -h -s 0 -N 1100000 -A 0.000.

ST type	Isolate identifier	Accession Number	Origin	Country
17	BL02	SAMN02360715	CLIN (Vitreous fluid)	USA
	C20	SAMN02360744	ENV	Unknown
	C23	SAMN02360745	ENV	Unknown
27	AZPAE14980	SAMN03105677	CLIN (Intra-abdominal)	USA
	BWHPSA011	SAMN02360683	CLIN (Tissue middle turbinate)	USA
	BWHPSA022	SAMN02360694	CLIN (Sputum)	USA
146	LES431	SAMN02641592	CLIN (Parent of CF patient)	UK
	LESB58	SAMEA1705916	CLIN (CF isolate)	UK
	AZPAE13757	SAMN03105416	CLIN (Respiratory Tract)	Canada
235	BTP032	SAMN03787333	CLIN	USA
	JJ692	SAMN02360667	CLIN (UTI)	USA
	NCGM2.S1	SAMD00061003	CLIN (UTI)	Japan
252	AZPAE12420	SAMN03105411	CLIN (CF isolate)	USA
	AZPAE15012	SAMN03105709	CLIN (Intra-abdominal)	Germany
	BWHPSA028	SAMN02360700	CLIN (Sputum)	USA
253	BL16	SAMN02360729	CLIN (Corneal scraping)	USA
	BWH058	SAMN02402442	CLIN	Unknown
	UCBPP-PA14	SAMN02603591	CLIN (Burn wound)	Unknown
274	AZPAE14926	SAMN03105624	CLIN (UTI)	Brazil
	AZPAE14981	SAMN03105678	CLIN (UTI)	France
	BWHPSA040	SAMN02360704	CLIN (Sputum)	USA
395	3581	SAMN02584694	CLIN	Unknown
	BWH059	SAMN02402443	CLIN	Unknown
	BWHPSA045	SAMN02360709	CLIN (Sputum)	USA

Table S2. Publicly-available genomes used for comparison with shared strains from the NCFB and CF cohorts. Where available, relevant information is provided on the origin of isolates. CLIN, Clinical; ENV, Environmental; UTI, Urinary Tract Infection. Based on the available information, there is no evidence that any of the isolates above are directly linked to our patient cohorts.

In silico prediction of hypermutators

Reads were quality and adapter trimmed using fastq-mcf using parameters -q 20 with skew settings switched off. Reads were aligned using bwa mem to the reference genome (*Pseudomonas aeruginosa* PAO1; NC_002516) and sorted and converted to BAM using samtools. The mpileup component of samtools was used to call variants and perform local re-alignment of sequences. The BCF formatted files were converted to VCF format using bcftools and filtered to exclude sites with coverage < 10 or variant quality < 60. The impact of variants was assessed using a custom perl script. Data was analysed using the Zeus computational infrastructure at the University of Exeter.

From the genome-wide list of variants, SNPs and insertion-deletion events (Indels) within seven genes implicated in proofreading and/or DNA repair were identified, namely *mutS* (PA3620), *mutL* (PA4946), *mutY* (PA0357), *mutM* (PA5147), *dnaQ* (PA1816), *mutT* (PA4400) and *uvrD* (PA5443). SNAP2 [8] and PROVEAN [9] were used to predict whether the observed SNPs and Indels would be neutral or deleterious with regards to protein function. There was complete concordance between the predictions from both methods. Predicted deleterious SNPs and Indels identified within *mutS*, *mutL*, *mutY*, *mutM* and *dnaQ* are presented in Table S3. No deleterious mutations were evident in *uvrD* or *mutT* in any of the sequenced isolates.

		Gene and nature of mutation				
Isolate[a]	**ST**[b]	***mutS***	***mutL***	***mutY***	***mutM***	***dnaQ***
PIB016	ST17	L52P	Q52X		L342P	
PIB026	ST17	Frameshift				
PIB045	ST27	V264E				
PIB001	ST235	ΔL541-S544				
PIB023	ST252					R33H
PIB058	ST252		H469R			R33H
PIB067	ST252					R33H
PIC030	ST252		Frameshift			R33H
PIB063	ST253	Frameshift		H72R		

Table S3. Prediction of hypermutators based on the identification of deleterious mutations in genes conferring DNA proof-reading and mismatch repair functions. All of the indicated amino acid substitutions are predicted to be deleterious by both SNAP2 and PROVEAN, whilst the frameshift mutations each cause premature truncation of the gene product. [a] PIB isolates are from NCFB patients, whilst PIC isolates are from CF patients. [b] Sequence Type, as defined by Multi-Locus Sequence Typing (MLST).

References

1. Connor TR, Loman NJ, Thompson S *et al*. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics*. 2016; 2
2. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30(15):2114-2120.
3. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014;15(3):R46
4. Bankevich A, Nurk S, Antipov D *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77.
5. Torsten, S. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30(14):2068-2069.
6. Winsor GL, Griffiths EJ, Lo R *et al.* Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic Acids Res*. 2016; 44(D1):D646-53.
7. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010; 11:595.
8. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. *BMC Genomics* 2015;16 Suppl 8:S1.
9. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015;31(16):2745-7.