Supplemental Material

Multi-omics approach identifies novel pathogen-derived prognostic biomarkers in patients with *Pseudomonas aeruginosa* bloodstream infection

Matthias Willmann[1,2], Stephan Göttig[3], Daniela Bezdan[4,5], Boris Maček[6], Ana Velic[6], Matthias Marschal[1], Wichard Vogel[7], Ingo Flesch[8], Uwe Markert[9], Annika Schmidt[1], Pierre Kübler[1], Maria Haug[1], Mumina Javed[1,2], Benedikt Jentzsch[1,2], Philipp Oberhettinger[1], Monika Schütz[1,2], Erwin Bohn[1,2], Michael Sonnabend[1,2], Kristina Klein[1,2], Ingo B Autenrieth[1,2], Stephan Ossowski[4,5,10], Sandra Schwarz[1], and Silke Peter[1,2]

[1]Institute of Medical Microbiology and Hygiene, University of Tübingen, Tübingen, Germany

[2]German Center for Infection Research (DZIF), partner site Tübingen, Tübingen, Germany

[3]Institute for Medical Microbiology and Infection Control, University Hospital, Goethe-University, Frankfurt am Main, Germany

[4]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[6]Proteome Center Tübingen, Auf der Morgenstelle, Tübingen, Germany

[7]Medical Center, Department of Hematology, Oncology, Immunology, Rheumatology & Pulmonology, University of Tübingen, Tübingen, Germany

[8]BG Trauma Center, University of Tübingen, Tübingen, Germany

[9]Clinic for General, Visceral and Vascular Surgery, Zollernalb Hospital, Albstadt, Germany

[10]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

Tables

Table S1: Basic prevalence characteristics of the patient study population (n = 166)

Parameter	Patients n (%)
Demographic characteristics	
Age ≥ 65 years	83 (50%)
Male sex	60 (36.1%)
Clinical characteristics	
Fatal outcome (30 days)	50 (30.1%)
Nosocomial infection	94 (65.6%)
Cardiac insufficiency*	24 (14.5%)
Myocardial infarction*	15 (9%)
Peripheral arterial disease*	18 (10.8%)
Cerebrovascular disease*	16 (9.6%)
Immunosuppression	110 (66.3%)
Chronic lung disease*	16 (9.6%)
Collagenosis*	10 (6%)
Liver disease*	11 (6.6%)
Diabetes*	49 (29.5%)
Renal disease*	22 (13.3%)
Malignancies	48 (28.9%)
HIV	1 (0.6%)
Surgery during hospitalization	66 (39.8%)
Infection-related characteristics	
Primary BSI	64 (38.6%)
Catheter-related BSI	17 (10.2%)
Genitourinary infection source	32 (19.3%)
Pulmonary infection source	32 (19.3%)
Intraabdominal infection source	17 (10.2%)
Wound infection source	21 (12.7%)
Concomitant infections	109 (65.7%)
Multiple pathogens in blood culture	32 (19.3%)

* As defined by Charlson Comorbidity Score [1]
BSI, blood stream infection; HIV, human immunodeficiency virus.

Table S2: Univariate analysis of patient-related and clinical factors in patients with *Pseudomonas aeruginosa* bloodstream infection

Parameter	Hazard ratio	95% CI	P-value
Age (years)†	1.005	0.9877 - 1.0226	0.56
Male sex	0.8	0.46 - 1.41	0.44
Nosocomial infection	1.25	0.7 - 2.24	0.44
Cardiac insufficiency*	1.24	0.6 - 2.55	0.57
Any cardiac comorbidity*	1.49	0.82 - 2.71	0.19
Peripheral arterial disease*	2.47	1.23 - 4.95	0.02
Cerebrovascular disease*	0.64	0.27 - 1.51	0.28
Immunosuppression	2.61	1.22 - 5.56	0.006
Steroids (>10 µg/day)	1.55	0.87 - 2.74	0.13
Neutropenia (during time at risk)	2.07	1.19 - 3.61	0.01
Diabetes*	1.22	0.68 - 2.19	0.51
Renal disease*	0.5	0.18 - 1.4	0.14
Malignancies	1.3	0.73 - 2.31	0.38
Charlson Score†	0.9731	0.8695 - 1.089	0.63
SAPSII (index day)†	1.0464	1.0279 - 1.0651	<0.001
Surgery during hospitalization	0.92	0.52 - 1.62	0.77
Primary BSI	1.44	0.81 - 2.53	0.22
Catheter-related BSI	0.46	0.14 - 1.48	0.14
Genitourinary infection source	0.15	0.04 - 0.63	0.0004
Pulmonary infection source	2.13	1.18 - 3.81	0.02
Intraabdominal infection source	0.63	0.23 - 1.74	0.34
Wound infection source	1.14	0.48 - 2.67	0.77
Concomitant infections	1.63	0.83 - 3.2	0.14
Multiple pathogens in blood culture	1.36	0.71 - 2.61	0.36
Creatinine (mg/dl) †	0.9551	0.73 - 1.2497	0.73
Appropriate antibiotic treatment	0.36	0.2 - 0.66	0.0021
Appropriate antibiotic treatment within 24 hours	0.67	0.38 - 1.17	0.17

† Continuous variable. The hazard ratio reflects the increase/decrease in mortality risk per unit increase.
* As defined by Charlson Comorbidity Score [1]
Blue-labeled variables were included in the clinical model (p-value < 0.2).
BSI, blood stream infection; 95% CI, 95% confidence interval; SAPSII, simplified acute physiology score II [2]

Table S3. Multivariate clinical model in patients with *P. aeruginosa* bloodstream infection

Parameter	Hazard ratio	95% CI	P-value
Immunosuppression	2.08	0.97 - 4.48	0.04
SAPSII (index day)*	1.0452	1.0272 - 1.0635	<0.001
Genitourinary infection source	0.19	0.05 - 0.8	0.0034
Appropriate antibiotic treatment	0.31	0.17 - 0.59	0.0007

* Continuous variable. The hazard ratio reflects the increase/decrease in mortality risk per unit increase.
95% CI, 95% confidence interval; SAPSII, simplified acute physiology score II [2].

The multivariate model illustrates the association of four clinical and patient-related factors with 30-day mortality. The model was regarded as a clinical model and was chosen as the basis for all screening models, where pathogen-derived factors were included one at a time.

Table S4. Association of accessory genome and core proteome cluster with 30-day mortality in patients with *P. aeruginosa* bloodstream infection

Parameter	Hazard ratio	95% CI	P-value	Group size (%)
Accessory genome clusters				
Acc-cluster 1	0.68	0.38 - 1.21	0.18	84 (50.6%)
Acc-cluster 2	1.95	1.005 - 3.79	0.06*	33 (19.88%)
Acc-cluster 3	1.44	0.69 - 3.03	0.35	29 (17.47%)
Acc-cluster 4	0.67	0.31 - 1.47	0.3	20 (12.05%)
Core proteome clusters				
Prot-cluster 1	1.24	0.63 - 2.42	0.54	39 (23.49%)
Prot-cluster 2	1.27	0.68 - 2.37	0.46	47 (28.31%)
Prot-cluster 3	0.99	0.53 - 1.84	0.97	36 (21.69%)
Prot-cluster 4	0.6	0.28 - 1.28	0.17	44 (26.51%)

*Wald-Test result: 0.048
95% CI, 95% confidence interval; Acc, accessory genome; Prot, proteome

Inclusion of accessory genome and core proteome clusters into the clinical Cox regression model revealed one accessory genome cluster (acc-cluster 2) associated with 30-day mortality, and thus identified as a high-risk cluster.

Table S5. Pathogen-derived prognostic biomarker candidates in the accessory genome gene and phenotypic screening models

Parameter	Hazard ratio	95% CI	P-value	Frequency (%)
helP	3.21	1.63 - 6.3	0.0021	22 (13.25%)
log-Prot7	2.68	1.36 - 5.27	0.0037	166 (100%)
log-Prot214	1.85	1.25 - 2.74	0.0012	166 (100%)
log-Prot330	0.16	0.05 - 0.55	0.0023	166 (100%)

95% CI, 95% confidence interval; Prot, protein
LFQ intensities were natural log-transformed for core proteomic data.

Table S5 shows hazard ratios and p-values of the four predictors after integration in the respective screening model. None of the tested variables had reached the Bonferroni corrected p-value threshold (p = 0.000021 for genomic variables, p = 0.000046 for protein level variables). However, the screening model was not a low complex association test. It had already incorporated clinical factors, such as physiological patient status (SAPS II score) and administration of appropriate treatment, which are known to be causally related to mortality. Thus, screening of genomic and protein level factors took already relevant confounders into consideration rather than being just a simple p-value evaluation and selection.

Table S6. Pathogen-derived factors included in the respective multivariate models

Dataset ID	Annotation
Accessory genome gene multivariate model	
gene372	TM2 domain protein
gene416	PAAR motif protein
gene457	hypothetical protein
gene686	hypothetical protein
gene1065	hypothetical protein
gene1087	Group II intron-encoded protein LtrA
gene1400	hypothetical protein
gene1799	hypothetical protein
gene1866	DEAD/DEAH box helicase (helP)
Phenotypic multivariate model	
corenewprot7log	Flagellar basal body protein FliL
corenewprot214log	Bacterioferritin
corenewprot330log	Putative aminotransferase
corenewprot1378log	Activator of ntr-like gene protein, OsmE
corenewprot4554log	Alkyl hydroperoxide reductase subunit F, AhpF
corenewprot5794log	Sulfate-binding protein Sbp
corenewprot6366log	Fumarate hydratase class II, FumC2
corenewprot8867log	Glutamine amidotransferase

Statistically significant pathogen-based factors ($p < 0.05$) with the 10% lowest p-values within the screening model were included in the multivariate models. Listed here are the nine factors for the multivariate accessory genome gene model and the eight factors for the multivariate phenotypic model.

Table S7. Functional annotation of prognostic biomarker candidates

Candidate	Annotation	UniProtKB	GO-Terms	GO-Identifier	Aspect
HelP	DEAD/DEAH box helicase	A6V9V7	DNA binding	GO:0003677	F
			helicase activity	GO:0004386	F
			ATP binding	GO:0005524	F
			hydrolase activity	GO:0016787	F
Prot7	FliL	A0A022NZV4	chemotaxis	GO:0006935	P
			bacterial-type flagellum dependent cell motility	GO:0071973	P
			bacterial-type flagellum basal body	GO:0009425	C
			membrane	GO:0016020	C
			integral component of membrane	GO:0016021	C
Prot214	Bacterioferritin	A0A069Q2A3	iron transport	GO:0006826	P
			cellular iron ion homeostasis	GO:0006879	P
			oxidation-reduction process	GO:0055114	P
			ferroxidase activity	GO:0004322	F
			ferric ion binding	GO:0008199	F
			oxidoreductase activity	GO:0016491	F
			cell	GO:0005623	C
Prot330	Probable aminotransferase	Q9HXJ9	biosynthetic process	GO:0009058	P
			catalytic activity	GO:0003824	F
			transaminase activity	GO:0008483	F
			transferase activity	GO:0016740	F
			pyridoxal phosphate binding	GO:0030170	F

P, process; F, function; c, component

Table S8. Putative virulence factors and non-synonymous SNPs

Putative virulence factor (core genome)	Annotation (locus tag)	Structural Alteration	SNP prevalence in dataset
alg8	Alginate biosynthesis protein Alg8 (PA3541)*	L408V	12%
algA	Phosphomannose isomerase / guanosine 5'-diphospho-D-mannose pyrophosphorylase (PA3551)*	V138L	12%
algB	Two-component response regulator AlgB (PA5483)*	L382R	34%
		T393A	81%
algC	Phosphomannomutase AlgC (PA5322)*	K313R	28%
algD	GDP-mannose 6-dehydrogenase AlgD (PA3560)*	-	-
algE	Alginate production outer membrane protein AlgE precursor (PA3544)*	G34N	43%
		V307I	13%
algF	Alginate o-acetyltransferase AlgF (PA3550)*	-	-
algG	Alginate-c5-mannuronan-epimerase AlgG (PA3545)*	-	-
algJ	Alginate o-acetyltransferase AlgJ (PA3549)*	-	-
algK	Alginate biosynthetic protein AlgK precursor (PA3543)*	A55T	11%
		A185T	12%
		A224V	58%
alg44	Alginate biosynthesis protein Alg44 (PA3542)*	-	-
algL	Poly(beta-d-mannuronate) lyase precursor AlgL (PA3547)*	-	-
algQ	Alginate regulatory protein AlgQ (PA5255)*	-	-
algR	Alginate biosynthesis regulatory protein AlgR (PA5261)*	-	-
algU	Sigma factor AlgU (PA0762)*	-	-
algX	Alginate biosynthesis protein AlgX (PA3546)*	E201D	16%
algZ	Alginate biosynthesis protein AlgZ/FimS (PA5262)*	-	-
alpR	Probable transcription regulator (PA0906)*	-	-
amrZ	Alginate and motility regulator Z (PA3385)*	-	-
aprA	Alkaline metalloproteinase precursor (PA1249)*	S113A	47.6%
		V335A	12%
		Q435K	20.5%
carR	Probable two-component response regulator (PA2657)*	-	-
carS	Probable two-component sensor (PA2656)*	P435S	16.3%
cdpR	Probable transcriptional regulator (PA2588)*	V84M	12%
		E154D	27.7%
cif	CFTR inhibitory factor, Cif (PA2934)*	S208A	87.9%
		D285E	47.6%
dksA	Suppressor protein DksA (PA4723)*	-	-
exoT	Exoenzyme T (PA0044)*	A83S	51.8%
		R163L	24.7%
		R302Q	13.3%
		G352S	51.8%
		Q360L	16.3%
exsA	Transcriptional regulator ExsA (PA1713)*	-	-
exsB	Exoenzyme S synthesis protein B (PA1712)*	R52G	32.5%
		Q105R	51.8%
		A127T	27.1%
fleQ	Transcriptional regulator FleQ (PA1097)*	-	-
fliO	Flagellar protein FliO (PA1445)*	-	-
gacA	Response regulator GacA (PA2586)*	-	-
icmF3	Hypothetical protein (PA2361)*	V14L	18.7%
		Y164F	12%
		D249N	30%
		E302D	14.5%
		A536G	10.8%
		A566G	30.8%
		D663N	66.3%
		T757S	28.9%
		K1045Q	22.9%
		R1081G	24%
		L1236Q	12%
		T1270I	10.8%
kinB	Probable two-component sensor (PA5484)*	Y50H	34.3%
		T74K	22.3%
		D112N	16.9%

		G141T	42.2%
		I466T	29.5%
lasA	LasA protease precursor (PA1871)*	L9M	66.3%
		P12S	21%
		A111V	59.6%
		E158G	12%
		G340S	36.7%
lasB	Elastase LasB (PA3724)*	Q102R	28.9%
		G241S	48.8%
lasR	Transcriptional regulator LasR (PA1430)*	-	-
lecA	LecA (PA2570)*	N89S	19.9%
morA	Motility regulator (PA4601)*	V88M	14.5%
		G98N	36.7%
		G124D	57.2%
		D495E	36.1%
mucA	Anti-sigma factor MucA (PA0763)*	-	-
mucB	Negative regulator for alginate biosynthesis MucB (PA0764)*	A211T	66.3%
mucC	Positive regulator for alginate biosynthesis MucC (PA0765)*	-	-
mucD	Serine protease MucD precursor (PA0766)*	I137V	50.6%
		Q225E	22.9%
		V441I	37.3%
muiA	Conserved hypothetical protein (PA1494)*	A70T	66.3%
		V266I	39.8%
		V307A	43.4%
		P329S	36.1%
		R373H	12%
		V511I	36.1%
mvaT	Transcriptional regulator MvaT, P16 subunit (PA4315)*	-	-
ndk	Nucleoside diphosphate kinase (PA3807)*	-	-
phzS	Flavin-containing monooxygenase (PA4217)*	Q154L	53%
		D256N	54.2%
plcH	Hemolytic phospholipase C precursor (PA0844)*	V39I	47%
		A212T	13.9%
		D338E	64.5%
		A390V	11.4%
		A530T	16.3%
		R665Q	23.5%
		L706F	19.3%
plcN	Nin-hemolytic phospholipase C precursor (PA3319)*	V342I	26.5%
popB	Translocator protein PopB (PA1708)*	A85V	52.4%
		R259K	41.6%
popD	Translocator outer membrane protein PopD precursor (PA1709)*	P39A	16.9%
		A57V	77.1%
		S101A	42.2%
		V193A	36.1%
		G245E	31.3%
pyrD	Dihydroorotate dehydrogenase (PA3050)*	K40E	12.7%
		K96R	61.4%
rpoN	RNA polymerase sigma-54 factor (PA4462)*	S76P	12.7%
rsmA	RsmA, regulator of secondary metabolites (PA0905)*	-	-
sbrI	Probable sigma-70 factor, ECF subfamily (PA2896)*	A7T	12%
tspR	Hypothetical protein (PA4857)*	-	-
eprS	Putative serine protease (PA14_18630)[#]	D4A	27.7%
		A35T	10.2%
		T216S	14.5%
		S504N	15.1%
		A634T	42.8%
		T653A	38.6%
		T710A	53%
		G717R	16.3%
		A812T	23.5%
		G909E	47%
		T917S	13.3%
feoA	Putative iron transport protein (PA14_56690) [#]	-	-
feoB	Putative ferrous iron transport protein B (PA14_56680) [#]	V48A	37.9%
		S165E	18.7%
		Q231R	32.5%

		L267M	12%
		V274I	23.5%
		V334A	13.9%
higA	Putative virulence-associated protein (PA14_61840) [#]	F23L	65%
		K88E	35.5%
		H100Q	15%

PAO1 (NC_002516.2)* and PA14 (NC_008463.1)[#] were used as reference templates for SNP calling. Strain ID186 (for PAO1 genes) and ID165 (for PA14 genes) were used as reference for the subsequent analysis in the regression models. All putative virulence factors that belong to the core genome dataset were investigated and listed here. If no structural variation was reported, there were no parsimony-informative SNPs identified that caused a replacement change and had a prevalence > 10% and < 90%.

Table S9. Helicase superfamily 2 members and *Pseudomonas species* RNA helicases

Figure label	Protein name	UniProt accession number	Organism
RNA helicases from Pseudomonas species			
G1	RL063	Q7WXZ7	*P. aeruginosa* PA14
G2	HelP	A6V9V7	*P. aeruginosa* PA7
G3	A9513_004725	A0A1B8THQ5	Pseudomonas sp. AU12215
G4	O164_02625	V7DI26	*P. taiwanensis* SJ9
DEAD-box helicases			
G5	SrmB	P21507	*E. coli*
G6	RhlB	P0A8J8	*E. coli*
G7	Ava_0642	Q3MFH0	*Anabaena variabilis*
G8	Ava_1952	Q3MBR2	*Anabaena variabilis*
G9	RhlE	P25888	*E. coli*
G10	CshE	Q81DF9	*B. cereus*
G11	CshC	Q81E85	*B. cereus*
G12	CshB	P54475	*B. cereus*
G13	CshA	P96614	*B. cereus*
G14	CshD	Q814I2	*B. cereus*
G15	DbpA	P21693	*E. coli*
G22	CsdA	Q46925	*E. coli*
RecQ			
G16	RecQ	P15043	*E. coli*
G17	RecQ	O34748	*B. subtilis*
Ski2-like			
G18	Sthe_0903	D1C273	*Sphaerobacter thermophilus*
G19	L687_03155	T5KDA6	*Microbacterium maritypicum*
G20	HR12_24870	A0A074TV47	Microbacterium sp.
G21	Mlut_11980	C5CBV6	*Micrococcus luteus*
DEAH-box helicases			
G23	HrpB	D9QDK8	*Corynebacterium pseudotuberculosis*
G24	RAM_06515	G0G7X3	*Amycolatopsis mediterranei*
G25	HprB	Q0I751	Synechococcus sp.
G26	HrpB	Q0C562	*Hyphomonas neptunium*
G27	HrpA	Q8NP89	*Corynebacterium glutamicum*
G28	HrpA	P43329	*E. coli*
G29	HrpA	O83538	*Treponema pallidum*

Figure labels indicate the labels in figure S6.

Table S10. Protein levels of secretion system effectors, type IV pilus system factors, and toxins in *helP*+ and *helP*- *P. aeruginosa* strains

Factor	Median LFQ intensities in *helP*+ strains (IQR)	Number of *helP*+ strains	Median LFQ intensities in *helP*- strains (IQR)	Number of *helP*- strains	p-value*
Type three secretion system					
ExoU	$7.54x10^6$ ($7.14x10^6$ - $4.35x10^7$)	5	$1.14x10^6$ ($5.41x10^5$ - $6.96x10^6$)	45	0.04
ExoS	$1.77x10^4$ (0 - $9.79x10^4$)	17	$1.35x10^4$ (0 - $5.65x10^4$)	96	0.82
ExoT	$3.85x10^4$ (0 - $1.08x10^5$)	22	0 (0 - $9.28x10^4$)	144	0.51
ExoY	0 (0 - $3.13x10^5$)	22	0 (0 - $2.39x10^5$)	131	0.71
ExsA	$1.83x10^6$ ($1.07x10^6$ - $5.97x10^6$)	22	$1.47x10^6$ (0 - $3.22x10^6$)	144	0.20
PopB	0 (0 - $2.61x10^5$)	22	0 (0 - $5.71x10^4$)	144	0.61
PopD	$6.29x10^4$ (0 - $1.34x10^6$)	22	0 (0 - $5.48x10^5$)	144	0.65
Type four pilus system					
PilQ	$7.65x10^5$ (0 - $1.2x10^7$)	21	$2.33x10^5$ (0 - $1.01x10^7$)	142	0.77
PilV	$7.05x10^6$ ($4.27x10^6$ - $1.07x10^7$)	22	$7.22x10^6$ ($4.56x10^6$ - $1.13x10^7$)	122	0.59
PilT	$3.47x10^7$($3.06x10^7$ - $4.32x10^7$)	22	$3.93x10^7$ ($3.12x10^7$ - $4.62x10^7$)	144	0.25
PilS	$3.24x10^6$ ($2.2x10^6$ - $1.45x10^7$)	22	$2.23x10^6$ (0 - $1.78x10^7$)	122	0.16
PilR	$2.2x10^7$ ($1.78x10^7$ - $3.06x10^7$)	22	$2.26x10^7$ ($1.67x10^7$ - $3.07x10^7$)	144	0.97
PilP	$3.23x10^7$ ($2.44x10^7$ - $4.66x10^7$)	22	$3.48x10^7$ ($2.58x10^7$ - $4.95x10^7$)	144	0.45
PilN	$3.06x10^7$ ($1.78x10^7$ - $3.26x10^7$)	21	$2.89x10^7$ ($2.2x10^7$ - $4.1x10^7$)	142	0.23
Exotoxins					
ToxA[†]	$5.65x10^4$ ($-6.10x10^4$ - $1.74x10^5$)	22	$9.65x10^4$ ($-7x10^3$ - $2x10^5$)	140	0.97

Factors were investigated when genes for the investigated factors were present in at least 50 strains. Only strains that had the gene for a respective factor were included in the analysis.

* by Mann-Whitney rank-sum test

[†] expressed in only 7 strains, mean (95% confidence interval) was used

LFQ, label free quantification units; IQR, interquartile range.

Table S10 shows differences in protein level expression of *P. aeruginosa* toxins or type IV pili system according to the presence of the *helP* gene. For each calculation, strains were only considered when they carried the respective gene of interest. For instance, only 50 strains carried the gene *exoU*, five of them were *helP* positive, 45 were *helP* negative. Protein levels of ExoU differed according to the *helP* status (p = 0.04), while all other factors showed no statistically significant differential protein expression according to the presence of the *helP* gene.

Table S11. Machine learning estimators of five distinct datasets using various classifiers and preprocessing strategies

Dataset	Machine learning classifier	Preprocessing strategy	ROC AUC	95% CI	fitting status
ACC	RF	None	0.781	0.22	overfitted
ACC	RF	PCA100	0.781	0.27	overfitted
ACC	RF	Percentile5	0.785	0.2	overfitted
ACC	SVM	None	0.698	0.2	fitted
ACC	SVM	PCA100	0.698	0.24	fitted
ACC	SVM	Percentile5	0.755	0.26	overfitted
ACC	LinSVM	None	0.826	0.17	overfitted
ACC	LinSVM	PCA100	0.691	0.22	overfitted
ACC	LinSVM	Percentile5	0.807	0.15	slightly overfitted
ACC	KNN	None	0.638	0.26	overfitted
ACC	KNN	PCA100	0.643	0.26	overfitted
ACC	KNN	Percentile5	0.678	0.36	overfitted
ACC	MLP	None	0.667	0.26	overfitted
ACC	MLP	PCA100	0.654	0.24	overfitted
ACC	MLP	Percentile5	0.751	0.22	slightly overfitted
Pheno	RF	None	0.711	0.25	overfitted
Pheno	RF	PCA100	0.718	0.28	overfitted
Pheno	RF	Percentile5	0.718	0.19	overfitted
Pheno	SVM	None	0.753	0.27	slightly overfitted
Pheno	SVM	PCA100	0.764	0.26	slightly overfitted
Pheno	SVM	Percentile5	0.8	0.2	fitted
Pheno	LinSVM	None	0.787	0.26	overfitted
Pheno	LinSVM	PCA100	0.729	0.24	overfitted
Pheno	LinSVM	Percentile5	0.817	0.21	slightly overfitted
Pheno	KNN	None	0.675	0.17	overfitted
Pheno	KNN	PCA100	0.634	0.2	overfitted
Pheno	KNN	Percentile5	0.693	0.3	overfitted
Pheno	MLP	None	0.71	0.27	overfitted
Pheno	MLP	PCA100	0.712	0.36	overfitted
Pheno	MLP	Percentile5	0.787	0.29	overfitted

SNP	RF	None	0.799	0.2	slightly overfitted
SNP	RF	PCA100	0.686	0.32	overfitted
SNP	RF	Percentile5	0.793	0.18	fitted
SNP	SVM	None	0.752	0.27	overfitted
SNP	SVM	PCA100	0.744	0.25	overfitted
SNP	SVM	Percentile5	0.748	0.23	fitted
SNP	LinSVM	None	0.793	0.18	overfitted
SNP	LinSVM	PCA100	0.79	0.25	overfitted
SNP	LinSVM	Percentile5	0.785	0.24	fitted
SNP	KNN	None	0.632	0.24	overfitted
SNP	KNN	PCA100	0.632	0.24	overfitted
SNP	KNN	Percentile5	0.752	0.22	slightly overfitted
SNP	MLP	None	0.768	0.23	overfitted
SNP	MLP	PCA100	0.748	0.25	overfitted
SNP	MLP	Percentile5	0.76	0.3	slightly overfitted
ALL	RF	None	0.725	0.2	overfitted
ALL	RF	PCA100	0.758	0.28	overfitted
ALL	RF	Percentile5	0.728	0.19	overfitted
ALL	SVM	None	0.689	0.22	overfitted
ALL	SVM	PCA100	0.667	0.24	overfitted
ALL	SVM	Percentile5	0.761	0.16	fitted
ALL	LinSVM	None	0.775	0.25	overfitted
ALL	LinSVM	PCA100	0.725	0.32	overfitted
ALL	LinSVM	Percentile5	0.79	0.24	overfitted
ALL	KNN	None	0.619	0.31	overfitted
ALL	KNN	PCA100	0.628	0.27	overfitted
ALL	KNN	Percentile5	0.647	0.38	overfitted
ALL	MLP	None	0.668	0.295	overfitted
ALL	MLP	PCA100	0.603	0.25	overfitted
ALL	MLP	Percentile5	0.743	0.25	slightly overfitted
Final	RF	None	0.788	0.26	slightly overfitted
Final	SVM	None	0.837	0.29	slightly overfitted
Final	LinSVM	None	0.829	0.33	fitted
Final	KNN	None	0.795	0.28	overfitted

Final	MLP	None	0.838	0.3	overfitted

* overfitting/underfitting assessment was done by a visual inspection of the learning curves

ACC, dataset with clinical risk factors and accessory genome features; Pheno, dataset with clinical risk factors, protein level and antibiotic susceptibility features; SNP, dataset with clinical risk factors and virulence gene variation features; ALL, dataset with clinical risk factors and accessory genome, protein level antibiotic susceptibility, and virulence gene variation features; Final, dataset with all features from the final Cox regression model; RF, random forest classifier; SVM, support vector machine classifier; LinSVM, support vector machine classifier; KNN, k nearest neighbor classifier; MPL, Multi-layer Perceptron; PCA100, transformation of all features in an array with a maximum of 100 components; Percentile5, feature array keeps only features that belong to the best 5% according to univariate feature/outcome relation based on p-value determination; ROC AUC; area under the receiver operating characteristic analysis curve; 95% CI; 95% confidence interval.

Blue labels indicate the best indicators from each dataset. These estimators were tested on the hold-out dataset (Table S12).

Table S12. Testing of machine learning estimators from each dataset on the hold-out dataset

Dataset	Machine learning classifier	Preprocessing strategy	ROC AUC	Matthews correlation coefficient	Correct positive classifi-cations (%)	Correct negative classifi-cations (%)	False negative classifi-cations (%)	False positive classifi-cations (%)
ACC	LinSVM	Percentile5	0.683	0.336	7	16	3	8
Pheno	SVM	Percentile5	0.595	0.197	4	19	6	5
SNP	RF	Percentile5	0.729	0.547	5	23	5	1
ALL	SVM	Percentile5	0.683	0.336	7	16	3	8
Final	LinSVM	None	0.895	0.726	10	19	0	5

Positive classification = estimator predicted risk of a fatal case
Negative classification = estimator did not predict risk of a fatal case

ACC, dataset with clinical risk factors and accessory genome features; Pheno, dataset with clinical risk factors, protein level and antibiotic susceptibility features; SNP, dataset with clinical risk factors and virulence gene variation features; ALL, dataset with clinical risk factors and accessory genome, protein level antibiotic susceptibility, and virulence gene variation features; Final, dataset with all features from the final Cox regression model; RF, random forest classifier; SVM, support vector machine classifier; LinSVM, support vector machine classifier; Percentile5, feature array keeps only features that belong to the best 5% according to univariate feature/outcome relation based on p-value determination; ROC AUC; area under the receiver operating characteristic analysis curve.

Figures

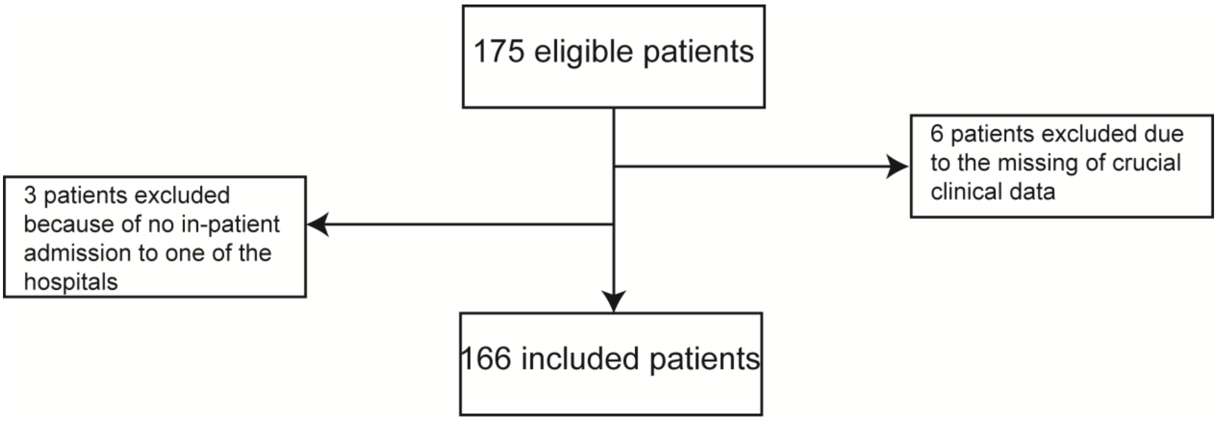
175 eligible patients
6 patients excluded due to the missing of crucial clinical data
3 patients excluded because of no in-patient admission to one of the hospitals
166 included patients

Figure S1. Flowchart of patient recruitment and exclusions

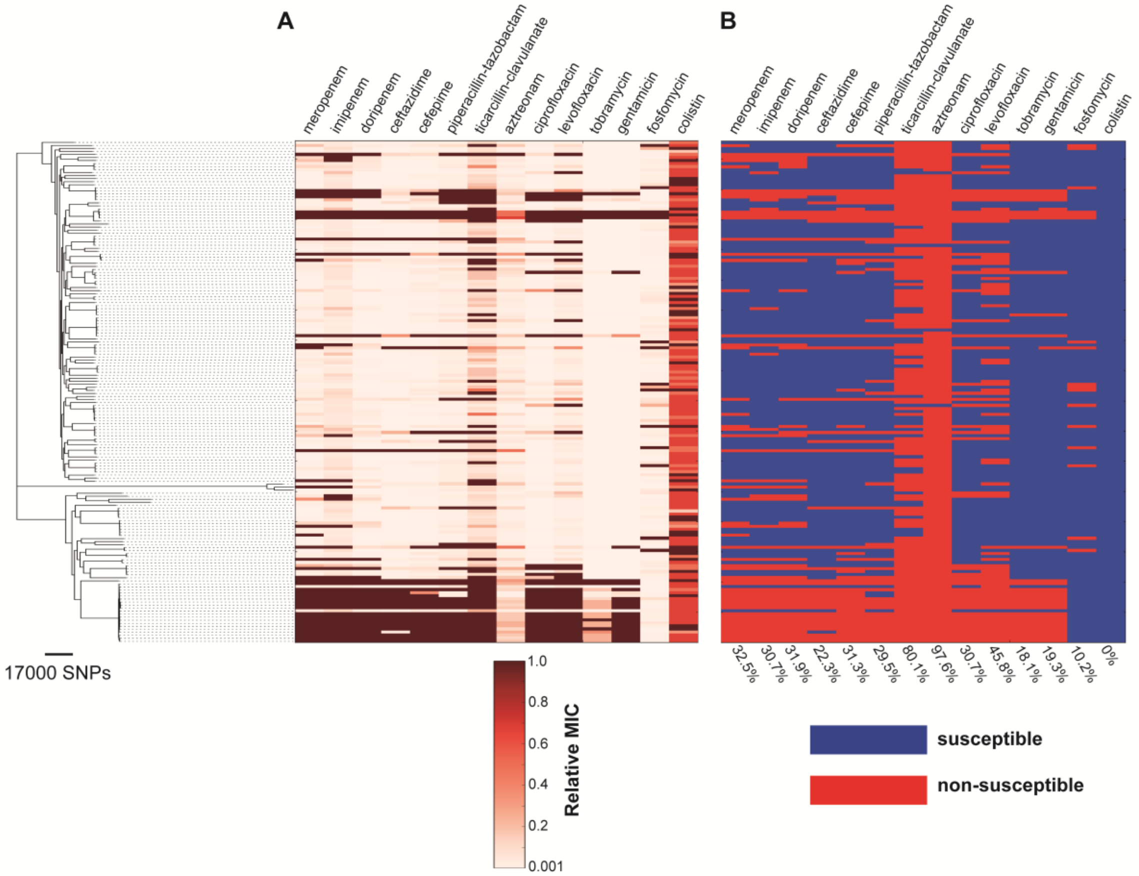

Figure S2. A) Distribution of relative minimum inhibitory concentration (MIC) values for various antibiotics relative to the maximum-likelihood tree. Relative MIC values are defined as ratio of the actual MIC value of a strain divided by the highest MIC value for the respective antibiotic in the dataset. The color bar below indicates the values. **B)** Distribution of antibiotic susceptibility for various antibiotics relative to the maximum-likelihood tree. EUCAST breakpoints were used for classification. Resistant and intermediate results were considered non-susceptible. Fosfomycin was rated susceptible when considered appropriate for combination therapy (epidemiological cut-off value 128 mg/L). Non-susceptibility rates are given for any antibiotic under the distribution map.

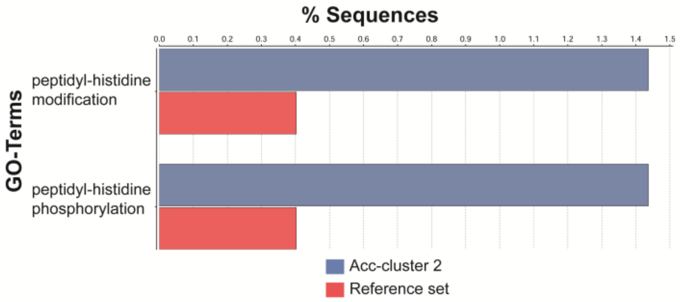

Figure S3. Enriched bar chart

Gene ontology (GO) terms from the acc-cluster 2 (13,943 gene clusters) were compared with the other three acc-clusters (reference set: 24,654 gene clusters) regarding a GO-term enrichment.

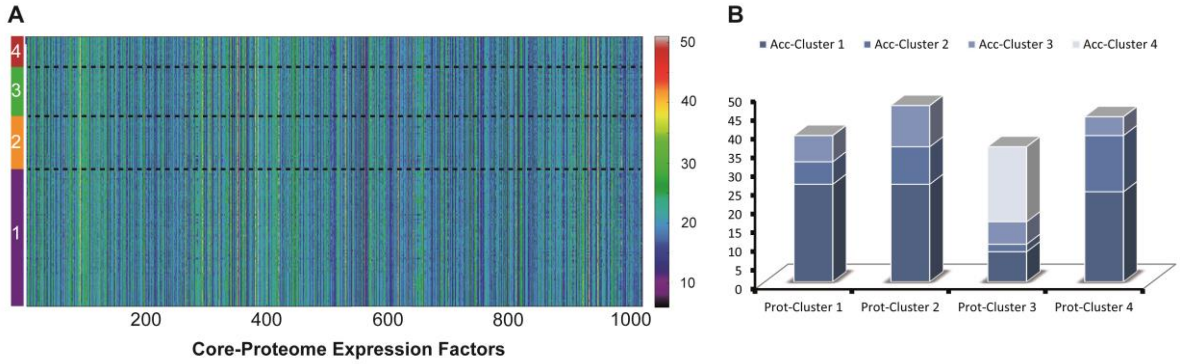

Figure S4. Core-proteome pattern analysis

(A) A heatmap of the root-normalized ($x^{1/6}$) protein level status from 1078 core proteins (x-axis) structured in blocks according to the accessory genome clusters (y-axis; acc-cluster 1 = purple, acc-cluster 2 = orange, acc-cluster 3 = green, acc-cluster 4 = red) revealed no pattern formation or significant distinction between blocks. A color bar on the right side displays the normalized protein level values. (B) The appearance of a protein level cluster in accessory genome clusters is displayed. It shows that strains from each core proteome (prot) - cluster derive from at least three different acc-clusters with the exception of acc-cluster 4. This genomic cluster was very distinct from the other acc-clusters (Fig 2A, red cluster) and produced only one protein level pattern (prot-cluster 3). However, this core proteome cluster was also a feature of the other three acc-clusters, thus still suggesting that there is not a strong relationship between the accessory genome and protein expression in *P. aeruginosa*.

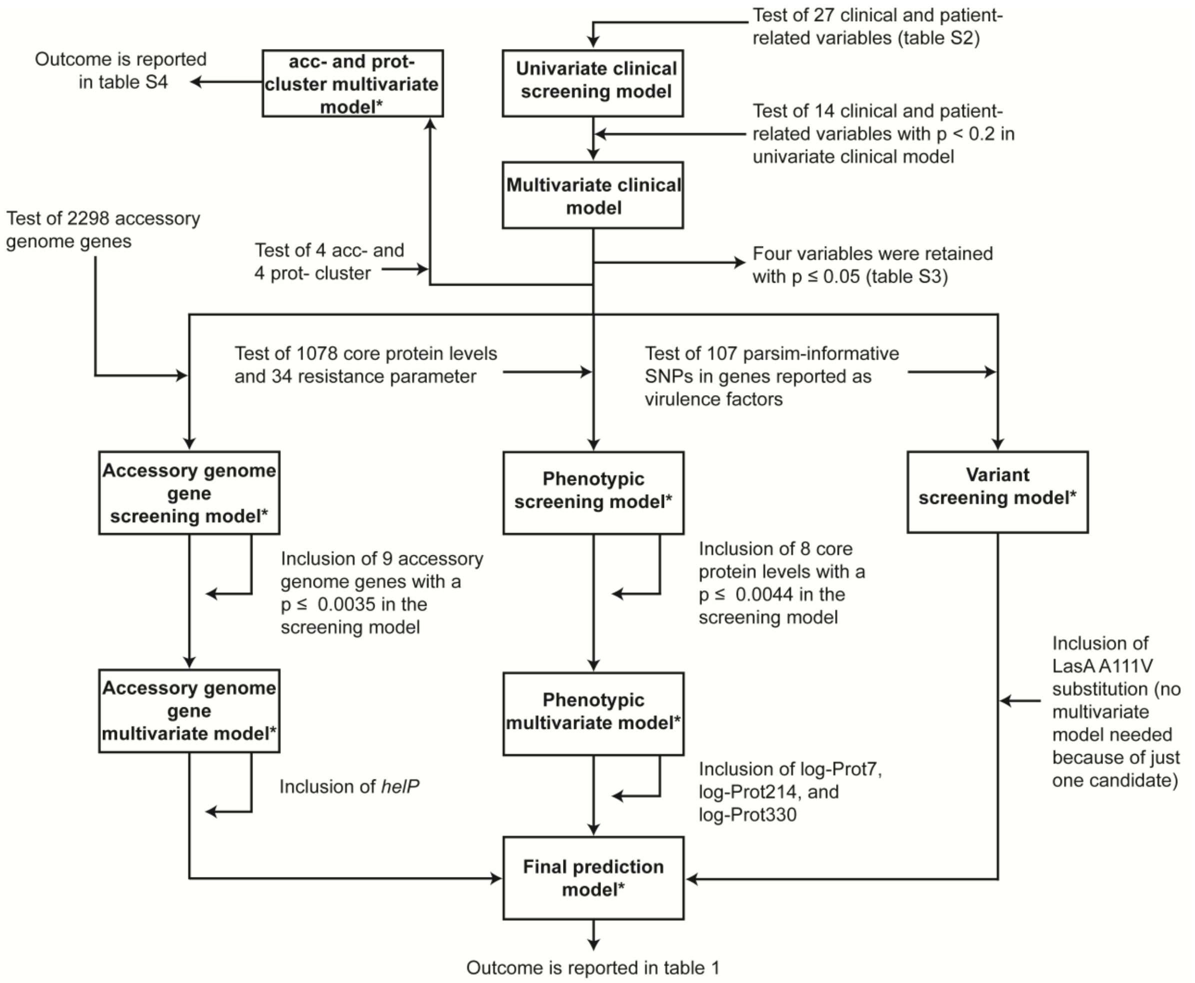
Test of 27 clinical and patient-related variables (table S2)
Univariate clinical screening model
Outcome is reported in table S4
acc- and prot-cluster multivariate model*
Test of 14 clinical and patient-related variables with p < 0.2 in univariate clinical model
Multivariate clinical model
Test of 2298 accessory genome genes
Test of 4 acc- and 4 prot- cluster
Four variables were retained with p ≤ 0.05 (table S3)
Test of 1078 core protein levels and 34 resistance parameter
Test of 107 parsim-informative SNPs in genes reported as virulence factors
Accessory genome gene screening model*
Phenotypic screening model*
Variant screening model*
Inclusion of 9 accessory genome genes with a p ≤ 0.0035 in the screening model
Inclusion of 8 core protein levels with a p ≤ 0.0044 in the screening model
Inclusion of LasA A111V substitution (no multivariate model needed because of just one candidate)
Accessory genome gene multivariate model*
Phenotypic multivariate model*
Inclusion of *helP*
Inclusion of log-Prot7, log-Prot214, and log-Prot330
Final prediction model*
Outcome is reported in table 1

Figure S5. Flowchart of the mortality predictor analysis

The four variables from the multivariate clinical model (table S3) were included in all models marked with an asterisk.

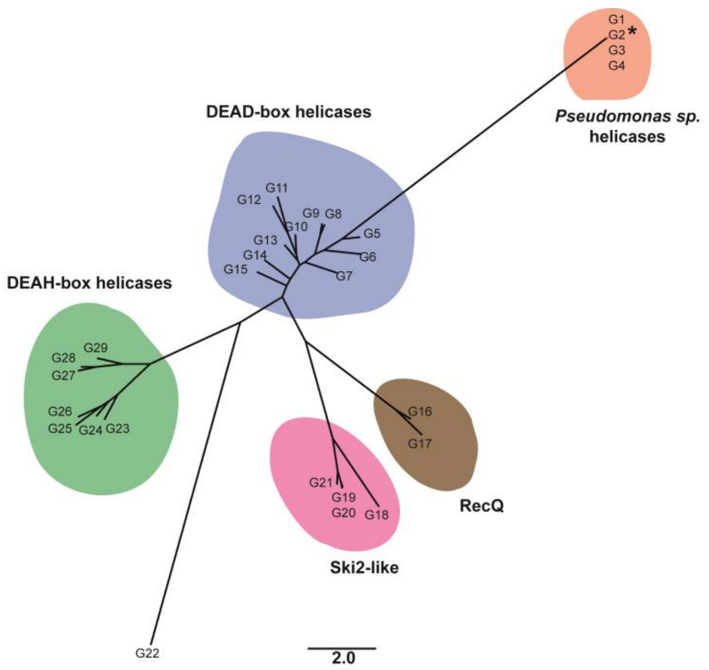

Figure S6. Unrooted tree of superfamily 2 of helicases

The maximum-likelihood tree is based on representative members of DEAD-box proteins, DEAH-box proteins, Ski2-like and RecQ-proteins. Annotations and UniProt accession numbers can be found in table S6. Proteins and phylogenetic methodology were chosen according to Redder et al [3]. The asterisk indicates HelP. The 29 protein sequences were aligned by ClustalW [4]. RAxML version (version 8.2.6) was deployed for tree reconstruction with 10000 bootstrap iterations using the "PROTGAMMAAUTO" command for best model determination [5]. FigTree (http://tree.bio.ed.ac.uk/software/figtree/) was deployed for tree visualization. The predicted helicases from *P. aeruginosa* were most closely related to DEAD-box helicases. Of note, G22 (CsdA from *Escherichia coli*), which is a presumed DEAD-box helicase, was phylogenetically distant from the rest of the DEAD-box helicases. The scale bar indicates the expected number of changes per site.

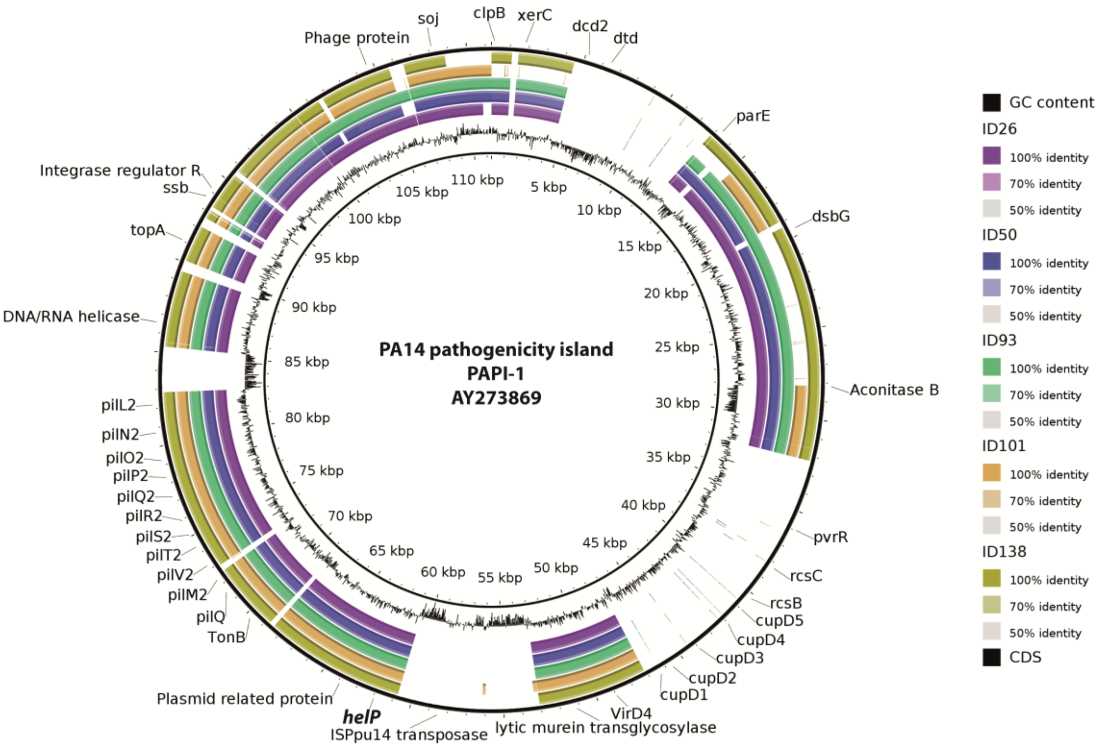

Figure S7. Reconstruction of the genomic environment of *helP*

Genomes of five *helP* positive strains were determined by PacBio long-read sequencing. Of these five, *helP* was predicted to be plasmid-encoded in four strains (ID26, ID93, ID101, and ID138) and located on the chromosome in strain ID50, based on Illumina sequencing data and plasmidSPAdes. PacBio sequencing determined a chromosomal location of *helP* in all strains. Since *helP* has a high similarity to a predicted DEAD/DEAH-box helicase (RL063) located on the PA14 pathogenicity island PAPI-1 (GenBank accession number: AY273869), we investigated whether *helP* was located within a similar genomic environment in our study strains. For this purpose, we performed a blastn comparison of a genomic stripe containing *helP* with PAPI-1 as reference, which is represented by the innermost ring, followed by a second ring that illustrated the GC content. The five following rings show the genomic environment of *helP* in all chosen strains (49 kb upstream and 61 kb downstream of *helP* according to the position of its homologous gene RL063 in PAPI-1). All strains have a genomic environment that highly resembles PAPI-1 with the exception of some downstream-located regions that are missing in all strains. Regional differences are also visible between the strains, indicating an independent evolution of the PAPI-1 related region. In all strains, a conjugative type IV pili apparatus of PAPI-1 is in close proximity upstream of *helP*. BRIG has been used for mapping visualization [6].

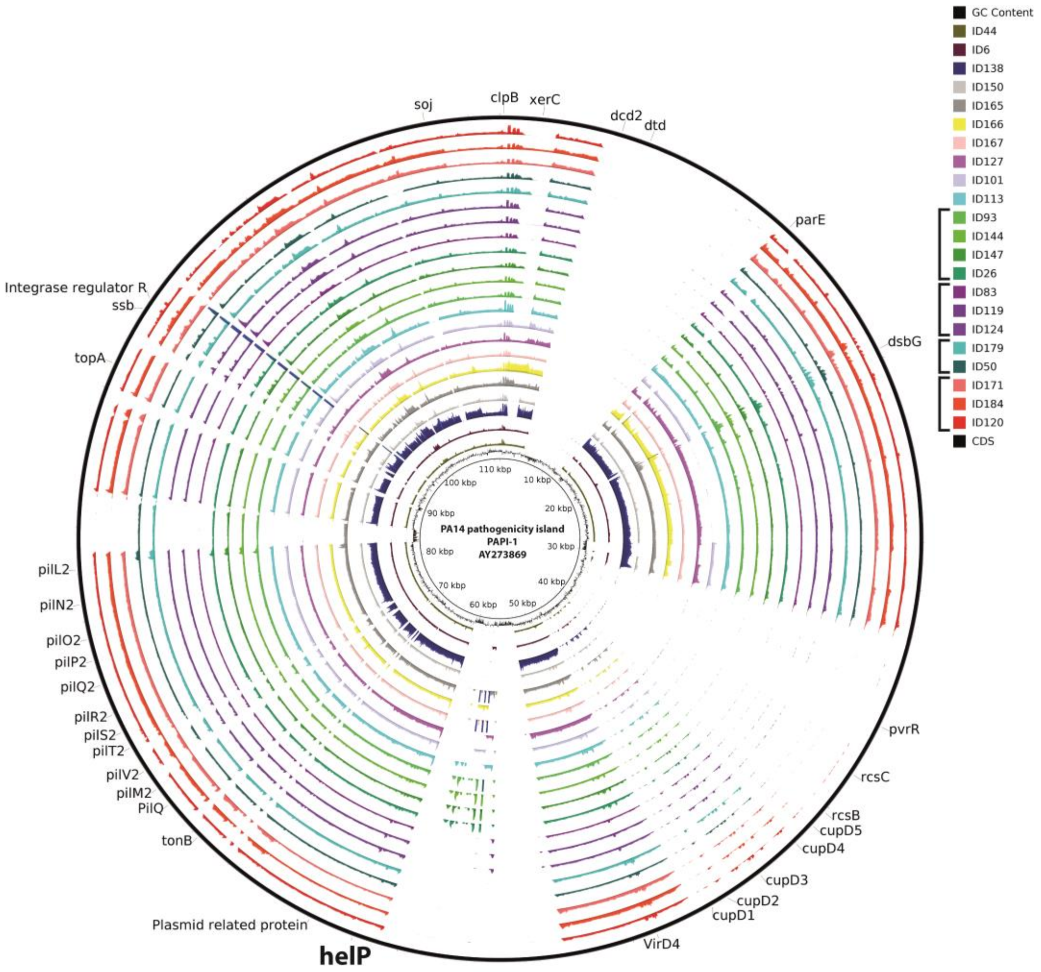

Figure S8. Reconstruction of PA14 pathogenicity island 1 (PAPI-1) components on the genome of *helP* positive *P. aeruginosa* strains

Sequence reads from the 22 *helP* positive *P. aeruginosa* strains were mapped against PAPI-1 (accession number: AY273869) using bwa-mem with a minimum mapping score of 30 [7]. Coverage and mapping visualization of sam files was performed using BRIG [6]. The innermost ring represents PAPI-1 reference, followed by a second ring that shows the GC content. All following rings demonstrate the coverage of sequencing reads from each *helP* positive strain over each position of the PAPI-1 reference (indicated by different colors, beginning with ID44 as first strain in the legend that refers to the third innermost ring and so forth). The ring’s height reflects the coverage depth. Strains originating from the same phylogenetic cluster are indicated by similar colors and are specifically labeled in the legend. There are 12 strains from four clusters and ten strains that do not genetically cluster with any other *helP* positive strain. In accordance with the detailed genomic environment analysis in figure S8, all strains contain large regions of PAPI-1, most likely in the same arrangement as the five strains in figure S8. The lack of the same regions indicates that these regions are also not present in more distant parts of the genome, with some minor exceptions in some strains. Generally, strains that form a phylogenetic cluster have a closely related coverage pattern, illustrating that the PAPI-1-related structure is conserved within a certain clone.

References

1. Charlson ME, Pompei P, Ales KL, MacKenzie CR: **A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.** *J Chronic Dis* 1987, **40:**373-383.
2. Le Gall JR, Lemeshow S, Saulnier F: **A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study.** *JAMA* 1993, **270:**2957-2963.
3. Redder P, Hausmann S, Khemici V, Yasrebi H, Linder P: **Bacterial versatility requires DEAD-box RNA helicases.** *FEMS Microbiol Rev* 2015, **39:**392-412.
4. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23:**2947-2948.
5. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.
6. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA: **BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons.** *BMC Genomics* 2011, **12:**402.
7. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.