Pseudomonas aeruginosa gene expression analysis using pangenome and PAO1 reference genomes.

BINP51, Bioinformatics: Master´s Degree Project, 45 credits

Department of Biology, Lund University, Lund, Sweden.

Student: Yi Su, email: yi3446su-s@student.lu.se

Supervisor: Magnus Paulsson, email: magnus.paulsson@med.lu.se

Abstract

Development in sequencing technologies has made the analyses of genetic material much more accessible. Processing sequenced data for an accurate analysis comes with its challenges, especially with the studies in microbial in clinical *in vivo* samples where difficulties in the collection of these samples for sequencing could lower the quality and contamination from the human host which might affect the accuracy of downstream analysis. In this project, we use RNA-seq and different reference genomes to look at the differential gene expression of *Pseudomonas aeruginosa* (PA), one of the most prevalent species of bacterial pathogens in the progression of chronic pulmonary diseases such as cystic fibrosis, due to its resistance to antimicrobial treatment. In this project, we created a pangenome from 21 strains of PA and explored the use of this, its subsets (core and soft-core gene sets) and a commonly used PA genome (PAO1) as reference genomes. We compared some of the differences and similarities in the results using the four gene sets, including for mapping transcripts while developing a feasible pipeline to process raw sample reads from human sputum samples for differential expression and gene ontology enrichment analysis. From the analyses, we have found differentially expressed genes upregulated in *in vivo* samples were related to biofilm, which plays a role in the difficulties in the treatment of PA infections, across the majority of the various genome reference-based results.

Introduction

Cystic fibrosis (CF) is an example of an autosomal recessive disease, which is inherited from mutations in the gene coding for cystic fibrosis transmembrane conductance regulator (CFTR) protein. These proteins reside on the surface of airway epithelial cells and the serous cells of the submucosal glands. Dysfunction or absence of CFTR leads to a complex complication of chloride absorption and sodium hyperabsorption which can result in obstructions of the airways. Accumulation of the airway surface liquid layer due to the thick and tenacious nature of the secretion hamper the ability to clear bacteria from the lower airways and thus, allowing the colonization of pathogens over time (Boucher, 2007).

The average estimated incidence of CF is between 1/3000 and 1/6000 births in the population of European descent. There are multiple individual factors which are associated with poor prognosis of CF, with lung function as the main predictor of survival. Other associated factors include female sex, higher age of diagnosis and early colonization of *Pseudomonas aeruginosa* (PA) (Scotet et al., 2020; Stephenson et al., 2017). Though bacterial infections may vary between clinics and countries, the pathogens PA and *Staphylococcus aureus* are most associated with CF (Uluer & Marty, 2014) .

Pseudomonas aeruginosa is a Gram-negative bacteria species which becomes more prevalent with the progression of pulmonary disease in adults with CF and remains the most important contributor to morbidity and mortality (Bhagirath et al., 2016). Once the bacterial colonization is established at an early age, PA can become complex and difficult to eradicate due to its genomic diversity and and adaptive resistance, despite high exposure to antibiotics (Rossi et al., 2020; Tai et al., 2015). The relatively large genome of PA and switching in gene expression allow the bacterial cells to survive challenges such as competition with other colonizers, antibiotics, osmotic stress, and host immunity, and adapt to the CF lung environment (Wu et al., 2014).

High-throughput sequencing technologies have been made much more accessible in recent years and a few studies have been deploying this to study differentially expressed genes in PA from *in vivo* clinical sputum samples and *in vitro* cultured isolates. Using a bioinformatics approach, sequences from these samples can be analyzed for gene expression using a pipeline of steps, resulting in a differential gene expression analysis in which the two environments are compared to each other. Although obtaining RNA sequences comes with their own complication that relates to sample collection, RNA extraction, library preparation and sequencing, there are still ways to improve the quality and processing time of their analysis.

The pangenome was first introduced by Tettelin et al., 2005 in the studies of multiple microbial strains, as the complete collective set of genes in the studied strains. Subsets from the pangenome include a core genome which is defined as the set of genes present in all strains, soft-core genome which includes genes that are present in most strains and an accessory genome which contains the collection of genes that are only present in only a few strains. Differential expression analysis presumes a common reference gene set to which the transcripts generated during sequencing can be counted. The result of this analysis will be highly dependent on the selected gene set, as only genes present in the selected reference gene set will be included in the analysis. Using a gene set that is too limited will cause loss of information and using an overly generous gene set will cause biases in the analysis as bacteria strains from the same species can carry very diverse genes in their genome.

To enable further studies in which the transcriptomic response of PA cells growing at two different environmental or clinical conditions, the bioinformatic analysis methods are important as they influence the results and biological interpretation. This project aims to develop a feasible pipeline and provide some insight into some of the different bioinformatic approaches and tools in RNA-seq analysis for PA from *in vivo* clinical samples. With the steps in the pipeline, we aim to pre-process raw read sequences from RNA-seq of clinical airway samples and deplete them of human reads. We will then create a pan-genome which includes core and soft-core genomes (Tettelin et al., 2005) using the 21 PA strains on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database with the tools Prokka (Seemann, 2014) and Roary (Page et al., n.d.). Along with these 3 genomes, the widely used PAO1 reference strain of PA will also be included in the analysis. We will then devise and test our approach by mapping transcripts to the reference pangenomes using the pseudo aligner kallisto (Bray et al., 2016); estimate the differential expression between publicly available transcripts from RNA-seq experiments (Cornforth et al., 2018) using sequences from the SRA Archive in R with the package DESeq2 (Love et al., 2014); conduct gene ontology analysis using PANTHER classification system (Thomas et al., 2003).

The established pipeline will be possible to use in further studies of bacterial pathogens in clinical airway samples compared to other environments, which may be relevant for detecting, understanding and controlling bacterial infections in the future.

Materials and Methods

Datasets

The complete genome assemblies and protein data used in the creation of the pangenome reference were downloaded from the NCBI genome database (Details in Supp table 1). *In vivo* and *in vitro* clinical sample data were downloaded from the NCBI Sequence Read Archive (SRA). PA isolates which were exposed to sub-MIC antimicrobials were chosen for the *in vitro* samples. The *in vivo* samples were clinical sputum samples from cystic fibrosis patients who were under antibiotic treatment. Accession numbers and details for the data used can be found in Supp table 2.

Pre-processing raw reads

Quality of the reads was assessed using FASTQC/version 0.12.1 (Andrew., 2010) to ensure the sequenced RNA data are viable to be used in the downstream analysis. Sequences were then trimmed with TrimGalore/version0.4.4 for adapter contamination. The raw RNA reads were collected and sequenced from the airways of clinical patients. As expected, there were large numbers of human reads in the sequence, which were removed before counting the reads. By removing human reads, we were able to process the files without the necessary security steps required when working with sensitive human data and reduce the file sizes for faster processing. The sample reads listed in the table was depleted of human reads using a combination of two different methods of detecting human reads: taxonomy classification method with the software Kraken2/version 2.1.1 (Wood et al., 2019) and alignment method software bowtie2/version 2.4.4 (Langmead & Salzberg, 2012).

The two-step method was used to ensure all human reads are removed from the *in vivo* samples. Kraken2 software was used for the first step in detecting human reads. Using the *.kraken* file outputs, the sequence ID for the reads that were not assigned by Kraken2 as '*Homo sapiens*' were saved as a list and used with seqtk/version1.2 subseq command to extract non-human reads from the sample reads files. The subsequent reads were then mapped to the human genome GRCh37 from NCBI using bowtie2/version2.4.4. SAM flags were interpreted using the Picard utility in the resulting SAM file output from bowtie2. SAMtools/version1.15.1 (Li et al., 2009) was used to find reads that were flagged as unmapped. These were then extracted into gzip compressed FASTQ files, completing the second step of removing human reads. A second Kraken2 report was made for the final cleaned product. Once this method was established, it was also applied in the decontamination of human reads in a parallel project focusing on the *in vivo* gene expression of *Haemophilus influenzae* (Polland et al., 2023).

PA pan-genome creation

The amount of plasticity in bacterial genomes creates a complication in the analysis based on their genetic material. Considering the different strains and variations, it is often difficult to find significant data with a reference genome based on one strain. Therefore, using a pangenome as a reference would potentially provide a more complete set of genes to explore.

Moreover, a core genome can be extracted from the pan-genome. The core genome was here defined as the set genes which were present in all 21 strains used to create the pangenome, and the soft-core genome is defined by the set of genes which were only present in 20 strains.

The pan-genome creation was delimited to the 21 strains of PA with complete assemblies of their genomes on the KEGG database. Genome assemblies were downloaded from NCBI and created into a reference pan-genome using the tools Prokka/version1.14.16 and Roary/version3.13.0. In the resulting pangenome, some genes could not be automatically assigned a locus tag based on the commonly used nomenclature for PA. Instead, these were labelled "*group_????",* which limits the possibility of downstream analyses. A custom Python script was generated to exchange "group_????" with NCBI locus tags using and the remaining sequences that were not reannotated were searched with DIAMOND/version2.1.4 (Buchfink et al., 2021) using the protein data from the 21 strains. From the final reannotated pangenome, the core and soft-core genomes were extracted and the pangenome was explored using the script provided with the Roary tool to generate statistics about the gene sets, a gene presence/absence matrix and phylogenetic trees.

Pseudo-alignment of sample reads

For this project, the pseudoaligner kallisto/version 0.48.0 was used in the pipeline to map the sample reads to the core, soft-core, pangenome and PAO1 reference with bootstrap value of

100, and the parameters for single-end sequence mapping were used. The resulting kallisto count data were used for downstream analysis.

Data normalization and exploration

Tximport/version1.28.0 was used to import kallisto count data into R/version4.3.1 language for statistical analysis. Counts were prefiltered where the genes with less than 10 counts across all samples were not included in the downstream analysis. Regularized logarithmic method rlog was chosen as the normalization method for visualization. The data was also explored with unsupervised clustering: PCA and hierarchical clustering, which provided a rough overview of the data before conducting differential expression analysis. This was also done to discover potential outliers in the samples that may askew any of the downstream analysis. R packages pheatmaps/version1.0.12 was used to create the heatmaps, using the default Euclidean method to create the sample-to-sample distance matrix.

Differential expression Analysis

The study design was set to compare the differentially expression genes between *in vivo* sputum samples from clinical patients under antibiotic treatment and *in vitro* lab-grown isolates which were also treated with antibiotics. The differential expression analysis was done in R with the use of R package DESeq2/version1.40.2. Significantly differentially expressed genes (DEGs) were considered as having an absolute log2 fold change > 1 and an adjusted p-value < 0.05. Locus tags of significant DEGs were searched on the Pseudomonas database pseudomonas.com.

Gene Ontology

Gene ontology (GO) classification of all genes in the core, soft-core and pan genome were explored using PANTHER release 17.0. Differentially expressed genes upregulated in the *in vivo* sputum samples for core, soft-core, pangenome and PAO1 reference-based results were analyzed with PANTHER Overrepresentation Test (Released 20230705) using the GO biological process annotation set, which also tested with Fisher’s Exact and correction for False Discovery Rate (FDR). Only the results with FDR p-value < 0.05 were included.

Results

Pangenome

The pangenome was created by the software tool Roary using the 21 strains of PA listed in the KEGG database and the annotations from Prokka. The core genome and soft-core genome were extracted from the pangenome and all three along with a PAO1 reference were used as reference for the pseudo-alignment of sample reads. The pangenome consisted of a total of 13065 sequences, while the core and soft-core genome subsets from it had 3144 and 1799 sequences respectively (Fig.1). Roary also defined the shell genome containing genes present between 19 and 3 strains and a cloud genome with genes present in 3 or less strains. The core genome consisted of the genes that were present across all 21 strains while the soft-core consists of the set of genes present in 20 strains, which were the number of strains for the different genomes, as indicated by Roary. A large proportion of the pangenome were genes that were only present in one or few of the strains used (Supp Fig.1).

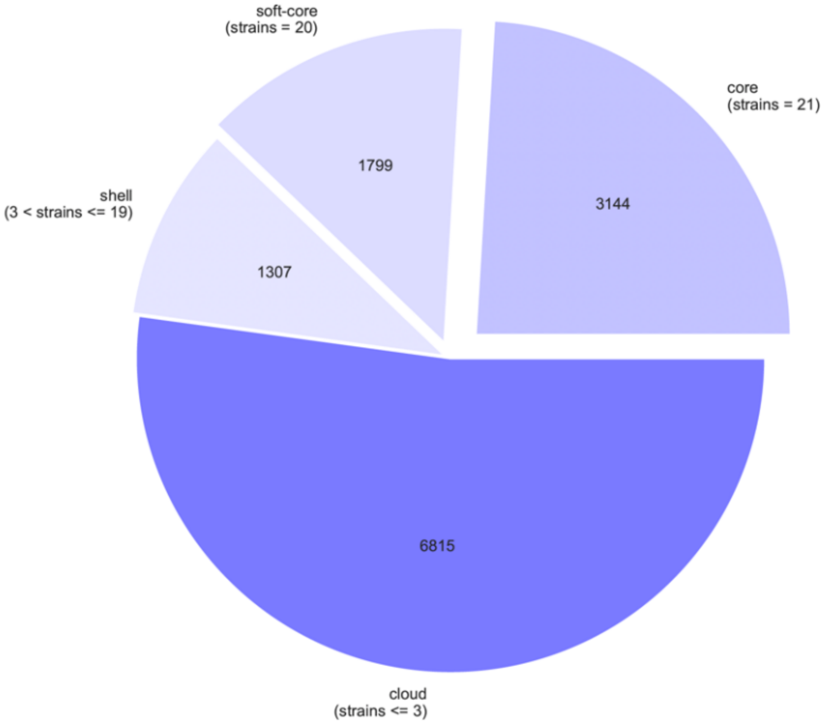


Fig.1 Pie chart illustrating the subsets of the pangenome generated from the 21 PA strains in the KEGG database. Outside the chart the subset is specified (core, soft score, shell and cloud genomes) and the number of strains that each the gene set is shared by. The number inside the chart denotes the number of genes included in each gene set.

The pangenome can also vary depending on the selection of strains used to create it. In the pangenome created in this project, PA7, a commonly used reference strain was included but there was also a large portion of genes that were not present in the core or soft-core genes (Fig.2). The inclusion or exclusion of such strains can have a significant effect on the number of resulting genes in different pangenome subsets.

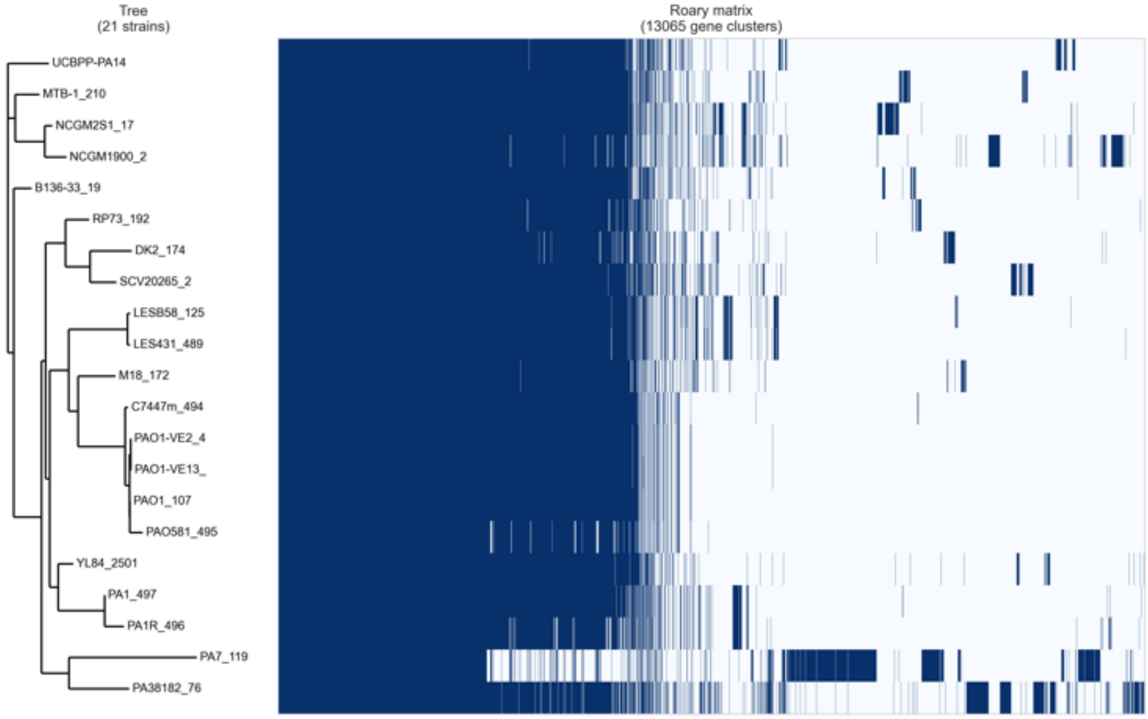


Fig.2 Matrix of the presence (blue) or absence of a gene (white) in the pangenome and a phylogenic tree of the 21 PA strains showing clustering of some strains such as the widely used reference strain PAO1_107 with other PAO1 strains, and a distinct pattern of gene absence/presence with strain PA7

To find out more about the predicted functions of the genes that were included in each gene set, Gene ontology (GO) classification of the genes present in the different genomes was explored under the PANTHER GO biological processes based on the locus tag of each gene; however, a large proportion of the genes were unclassified by PANTHER. The percentage of unclassified genes in each total number of genes in the 3 different genomes was 59.9% for pangenome, 53.6% for core and 68.2% and for soft-core genomes. Most of the genes that were classified by PANTHER have the GO category for cellular process and metabolic process for all 3 genomes while the core genome had a higher percentage of these genes in its genome compared to the pangenome and the soft-core genome. In contrast, genes under the GO category biological adhesion were completely absent from the core genome gene set. No genes from the soft-core genome (set of genes present in 20 strains) were categorized under the terms “reproduction and reproductive process”. Higher percentages of genes for GO categories were seen with the core which was likely due to the lower percentage of unclassified genes in its genome compared to the other two gene sets (Fig.3A). To account for this, the results for unclassified genes were filtered out, and the proportion of each GO category in the total categorized genes was calculated and plotted in Fig3.B to present the differences between reference genomes more accurately. Without including the unclassified genes, the proportion of genes under each GO category between all 3 references was quite similar, except for biological regulation which was lower in the core compared with the soft-core and pangenome (Fig.3B).

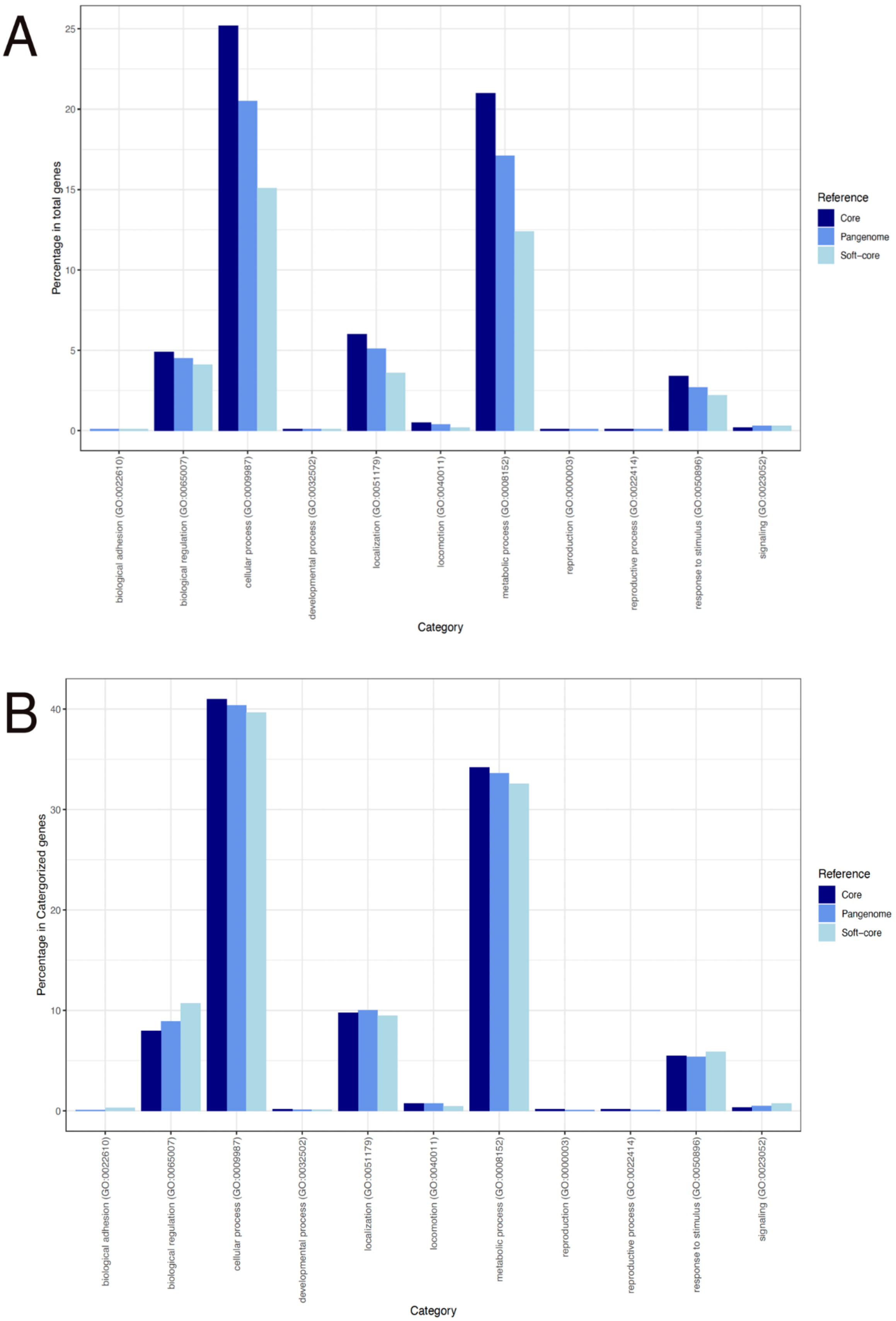

A
Percentage in total genes
25
20
15
10
5
0
biological adhesion (GO:0022610)
biological regulation (GO:0065007)
cellular process (GO:0009987)
developmental process (GO:0032502)
localization (GO:0051179)
locomotion (GO:0040011)
metabolic process (GO:0008152)
reproduction (GO:0000003)
reproductive process (GO:0022414)
response to stimulus (GO:0050896)
signaling (GO:0023052)
Category
Reference
Core
Pangenome
Soft–core
B
Percentage in Catergorized genes
40
30
20
10
0
biological adhesion (GO:0022610)
biological regulation (GO:0065007)
cellular process (GO:0009987)
developmental process (GO:0032502)
localization (GO:0051179)
locomotion (GO:0040011)
metabolic process (GO:0008152)
reproduction (GO:0000003)
reproductive process (GO:0022414)
response to stimulus (GO:0050896)
signaling (GO:0023052)
Category
Reference
Core
Pangenome
Soft–core

Fig.3 Percentage of all genes (A) and genes given a GO-term (B) of the total number of genes in each gene set under the GO term for biological processes. Majority of genes can be seen among the categories cellular process and metabolic process, as well as localization, biological regulation, response to stimulus categories for all 3 genomes.

Pre-processing PA sample reads

To test the created bioinformatic pipeline, sample sequence reads downloaded from the SRA archive and were trimmed for adapter contamination before further decontamination processing. *In vivo* human sputum sample sequences naturally contained human reads, and these reads were depleted from the microbial sequences. The sequences were filtered twice, first with Kraken2 then with bowtie2 to detect human reads, and no human reads were detected in the resulting sequences by a second kraken report after the two filtering. For some of the samples, a substantial percentage (approx. 60%) of the total reads remained after the human reads removal process while most samples only have about 11-38% of their total reads remaining. (Table 1)

Table 1 Number of reads for in vivo sputum samples for human reads decontamination process.

Sample/ SRA Accession	**No. of reads**						
	Before human reads removal				*After human reads removal*		
	Total reads	*Classified as human - Kraken2*	*Total remaining reads after 1st removal*	*Remaining reads classfied as human - bowtie2*	*Total reads*	*Total (%) that remain*	*PA reads - Kraken2*
SRR6833347	89739225	29153461	60585764	7711681	52874083	58,92	37209
SRR6833344	53273099	42045528	11227571	5277978	5949593	11,17	240249
SRR6833345	80472332	55501895	24970437	15698029	9272408	11,52	226354
SRR6833346	40794451	18438229	22356222	13031965	9324257	22,86	30866
SRR6833349	70634441	26181099	44453342	17182543	27270799	38,61	1134684
SRR6833350	35200463	8902173	26298290	18731251	7567039	21,50	12325
SRR6833351	20062069	2550654	17511415	4968907	12542508	62,52	26775

Differential expression analysis

The count data output files from the kallisto pseudoalignment were imported into R using tximport. 4 types of reference genomes were used for the alignment: core, soft-core, pan and PAO1 reference genome, steps in the analyses were repeated for each category and the results were compared between them.

The raw counts were normalized with variance-stabilizing transformation (VST) and regularized logarithmic method, then plotted the standard deviation of the transformed data against the mean. The rlog method was chosen as the normalization method for visualization over VST since the standard deviation was seen as more constant for all 4 sets of count data (Supp fig.2).

For the exploration and visualization of the imported data, principal component analyses (PCA) were performed for all 4 categories: core, soft-core, pan and PAO1 reference genome pseudo-aligned count data. It can be inferred from the PCA plots that in vitro samples cluster closer together than the *in vivo* sputum samples, and this is consistent in all 4 reference genome categories. *In vivo* samples are expected to have high variability since there can be a lot of contributing factors in differences in patients' co-morbidities, antibiotic treatment, genetics etc compared to a controlled laboratory environment. One of the sputum samples was more distant from the group, possibly due to low sequence quality or coverage.

Hierarchical clustering of samples visualized in heatmaps also shows the two groups, *in vivo* and *in vitro* grouping together and the same *in vivo* sample clustering further from the *in vivo* group, but still relatively closer compared to *in vitro* sample group. This pattern was also seen in the data from other genome references.

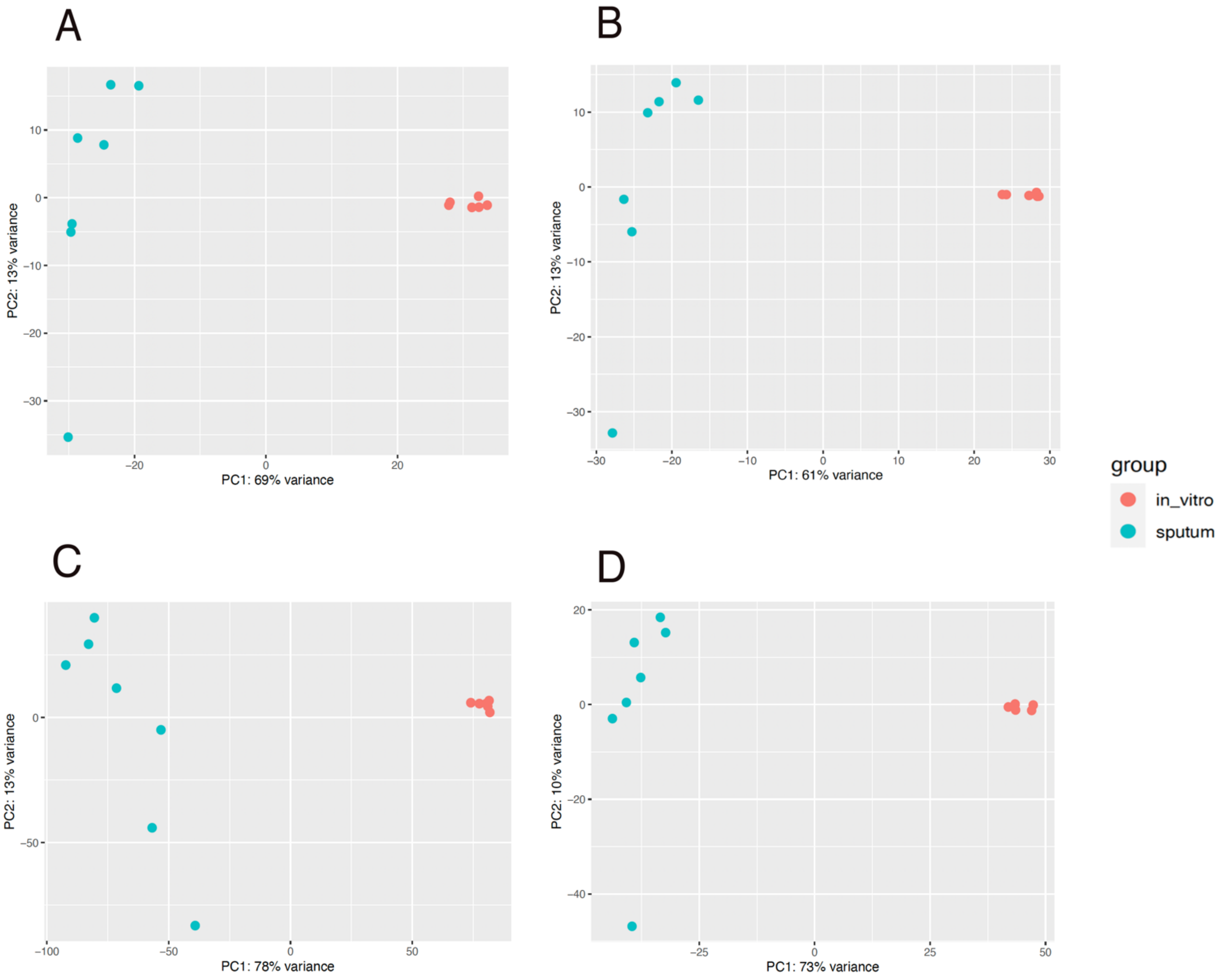


Fig. 4 Principal component analysis (PCA) displaying PC1 and PC2 of samples for core (A), soft-core (B), pan (C) genome and PAO1 reference genome (D) showing that *in vitro* samples group together closely, compared to the *in vivo* sputum samples that were further apart.

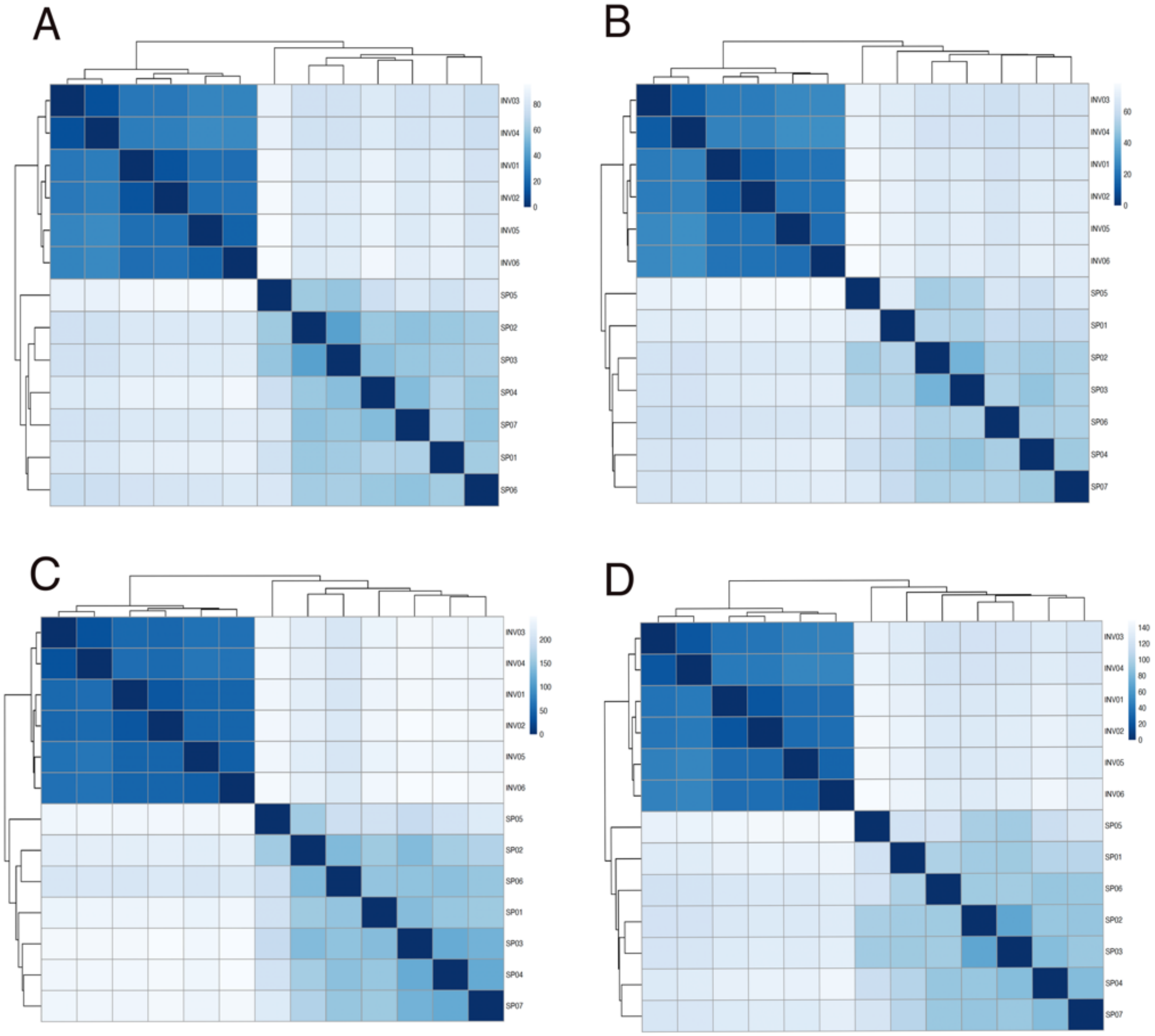


Fig. 5 Sample distance heatmap of in vivo and in vitro counts in using core (A), soft-core (B), pan (C) and PAO1 reference genomes. Clearer visualization of one of the sputum samples clustered further away but still closer with samples in its group.

Using the different genomes, the number of significantly differentially expressed genes are 1452 for core, 838 for soft-core, 3860 with pangenome and 2705 with PAO1 reference. Significantly differentially expressed genes are defined as having an absolute value of log2 fold change (LFC) > 1 and p-adjusted value > 0.05 (table).

From the core genome reference counts, alginate biosynthesis related genes and other genes involved in biofilm formation are more prevalent with the highest LFC. These alginate biosynthesis related genes were not present in the other 3 datasets. Some of the upregulated soft-core DEGs can be found in the psl cluster (locus PA2231-PA2245) involved in psl polysaccharide synthesis, which is important in the biofilm structure of PA (Wei & Ma, 2013). These psl locus tags, such as PA2231, *pslA* were commonly found in the soft-core and PAO1 reference DEGs. PA4107, a stress response and virulence modulator under high Calcium concentration, and PA4101 biofilm maturation regulator (Fan et al., 2021; Sarkisova et al., 2014) were the top DEGs in highest LFC for PAO1. Overall, more hypothetical proteins are found in the soft-core, pangenome and PAO1 reference compared to the core (Supp tables 3.1-3.4).

Table 2 Top 10 significantly differentially expressed genes upregulated with highest LFC in *in vivo* sputum samples (with core genome reference)

Locus	Product	LFC	padj
PA3546	alginate biosynthesis protein AlgX	8.73	1.81e-15
PA3540	GDP-mannose 6-dehydrogenase	8.72	2.42e-28
PA3557	4-amino-4-deoxy-L-arabinose-phospho-UDP flippase subunit E	7.87	2.10e-15
PA4495	hypothetical protein	7.75	1.54e-34
PA3551	mannose-1-phosphate guanylyltransferase	7.68	3.06e-25
PA4883	hypothetical protein	7.64	8.49e-48
PA3601	50S ribosomal protein L31	7.54	1.49e-123
PA3541	glycosyl transferase	7.52	1.13e-27
PA4836	hypothetical protein	7.47	2.33e-41
PA3544	alginate biosynthesis protein AlgE	7.29	3.75e-20

Gene Ontology Enrichment analysis

Upregulated DEGS in the *in vivo* samples for all 4 reference-based results were analyzed for GO terms in biological processes using PANTHER. Only 3 GO terms were found with the upregulated DEGs for sputum samples in the core dataset, these terms were overrepresented and categorized under alginic acid metabolic process and monoatomic ion transport (Table 3). Interestingly, in both the results for the pan genome and PAO1 reference, GO terms related to iron transport were overrepresented while much more genes were underrepresented, including multiple terms related to metabolic processes and cellular biosynthetic processes. The majority of the GO terms for the soft-core, however, were unclassified by PANTHER and these were overrepresented, while several terms associated with metabolic and biosynthetic processes were, similarly with pan and PAO1, underrepresented (Table 4). All significantly overrepresented and underrepresented terms for each were stored in tables. (For a detailed list of GO, see Supp table 4.1-4.4.)

Table 3 GO enriched terms for upregulated DEGS in *in vivo* samples (core)

GO biological process complete	**PA-REFLIST (5564)**	**Count (882)**	**Expected**	**Over/Under represented (+/-)**	**Fold Enrichment**	**Raw P-value**	**FDR**
alginic acid metabolic process (GO:0042120)	16	13	2.54	+	5.13	4.39E-05	9.76E-02
monoatomic cation transport (GO:0006812)	67	28	10.62	+	2.64	5.48E-05	4.07E-02
monoatomic ion transport (GO:0006811)	71	29	11.25	+	2.58	4.95E-05	5.51E-02

Table 4 Top 5 GO terms with lowest raw p-values for upregulated DEGs in *in vivo* results

	GO biological processes	**Over(+)/ Under (-) represented**
Core	alginic acid metabolic process (GO:0042120)	+
	monoatomic ion transport (GO:0006811)	+
	monoatomic cation transport (GO:0006812)	+
	-	
	-	

Soft-core	cellular process (GO:0009987)	-
	macromolecule metabolic process (GO:0043170)	-
	Unclassified (UNCLASSIFIED)	+
	biological_process (GO:0008150)	-
	primary metabolic process (GO:0044238)	-
Pan	cellular nitrogen compound metabolic process (GO:0034641)	-
	nucleobase-containing compound metabolic process (GO:0006139)	-
	nitrogen compound metabolic process (GO:0006807)	-
	organonitrogen compound biosynthetic process (GO:1901566)	-
	translation (GO:0006412)	-
PAO1	cellular nitrogen compound metabolic process (GO:0034641)	-
	nucleobase-containing compound metabolic process (GO:0006139)	-
	nucleic acid metabolic process (GO:0090304)	-
	nitrogen compound metabolic process (GO:0006807)	-
	gene expression (GO:0010467)	-

Discussion

To have a better insight into the difference in gene expression of bacteria between *in vivo* sputum samples of cystic fibrosis patients and cultured bacteria under controlled environments when treated with antibiotics, we used an RNA-seq pipeline using different reference genomes for transcript mapping, to analyze *Pseudomonas aeruginosa*, one of the prevalent bacterial species in progressive pulmonary disease patients which show antimicrobial resistance towards treatment.

A pangenome of 21 different PA strains was created for use as reference in the pseudoalignment of the RNAs-seq samples. The PA pangenome is an open pangenome where the number of genes continuously grow exponentially with new strains added. Therefore, to create a pangenome that was feasible for the resources available for this project, only 21 strains with complete genomes and annotations on the KEGG database were used, which include the most well-studied strains. The choice of using a pangenome was considered because of the nature of bacterial genetic material and with the aim to include genes shared by some of the different strains that might not be present in a reference strain like PAO1. Genes in the core genome are most likely to be comprised of more well annotated conservative genes for maintaining their biological processes, therefore the soft-core genes were included in this project since antimicrobial resistance and virulence genes might vary from strain to strain.

Obtaining RNA-seq data from in vivo samples can be challenging with the high standards required for the extraction process of the genetic material in question, and the variation between the samples can prove difficult for any downstream analysis that depends highly on the quality of the sequences. The initial plan for this project involved using our own RNA sequences, however, due to the low quality of these sequences in the samples, RNA sequences from the SRA database were used instead. To improve the quality of the analysis, various bioinformatic approaches have been developed and employed to process the data, including trimming and human reads removal. Aggressive trimming of sequences can have a significant effect on gene expression analysis especially on short reads sequences (Williams et al., 2016). The *in vivo* sample sequences also had human reads naturally since the samples were sputum samples collected from clinical patients. To decontaminate human reads from microbial reads for downstream analysis and faster processing, the *in vivo* sample sequences were filtered. A previous study showed that using two-step methods to remove human reads, produced some of the better results in decontaminating microbial samples and different methods of detecting human reads in microbial sequencing datasets have been tested by (Bush et al., 2020). Detecting and removing human reads also has potential consequences. If certain genes in the bacteria have high similarity with genes that are classified as human, in such a case this could potentially cause a loss of data in the differential expression analysis between *in vivo* and *in vitro* samples. Kraken2 was used to detect the final decontaminated sequences, in future studies, the use of a third software tool or database may be recommended to confirm instead.

The tool kallisto pseudo-aligns transcripts to an annotated reference and includes the quantification step of the counts, producing a raw count matrix which can then be imported into DEG tools for analysis. The pseudo-alignment by kallisto does not require high computing power, is much faster and the memory usage is low enough to be used on a personal laptop. Some traditional aligners provide more data on the mapping, such as a splice junction aware STAR and a quantification step would be required to produce count data. Since these details are not required in the downstream analysis, a more lightweight tool like kallisto was used. A previous comparison study of different alignment tools also showed that another pseudoalignment tool, salmon, would provide similar results (Schaarschmidt et al., 2020). Normalization of raw counts is a staple for differential expression analysis and various methods or approaches to normalization exists and their use depends on the nature of the data at hand. In this project pipeline, the R package DESeq2 was used. The data was tested using different transforms and regularized logarithmic method showed the most constant standard deviation across all 4 sets of count data and has been shown to be generally performed well against other methods (Love et al., 2014).

Among the top upregulated DEGs *in vivo* sputum samples, many were involved in biofilm. Although, using different reference genomes found DEGs related to alginate synthesis biofilm formation was prevalent in core, likewise for biofilm structure related in soft-core and PAO1 reference, and maturation related genes in PAO1. These genes that are involved in the biofilm can be found in mucoid-type PA strains of cystic fibrosis patients and poses a difficulty in their treatment due to antimicrobial resistance. Gene ontology enrichment analysis also showed different results with different genome references. GO terms in metabolic and cellular processes were underrepresented in pan and PAO1 genomes, alginic acid metabolism was overrepresented in core and an overrepresentation of unclassified in soft-core. It makes sense that vital functions for cell replication and metabolism are shared between all strains and are part of the core genome. In future studies using a pangenome, the soft-core genome would desirably include the core genome as well to provide more insight, since the definition for a soft-core can be more flexible than only having a set of genes that were shared between 20 strains in this current project. Although only upregulated DEGs in *in vivo* samples were analysed in this project, it would be important and recommended to also include downregulated DEGs that potentially show the contrast between the cultured PA and *in vivo* samples.

In conclusion, the choice of reference genomes to which the transcripts were pseudo-aligned resulted in different DEGs upregulated in *in vivo* samples and with GO terms in biological processes. There were distinct DEGs found in the core and soft-core datasets that may prove insightful into reasons for antibiotic resistance due to biofilm or virulence. Results from the pangenome and PAO1 reference showed similar GO terms in this project, it may be inferred that using PAO1 reference would suffice if using a pangenome is not feasible. In future pangenome studies, a core genome or an expanded soft-core genome may be used to discover a more specific set of genes.

Transcriptomic analysis of PA in *in vivo* clinical samples can be a challenge, however there are bioinformatic approaches where the quality of the analysis can be improved. Exploring the options in using a pangenome compared to a single reference genome provided more insight into the classfications of genes that may be expressed with using each different genome as a reference for the mapping sample sequences. Quality control should be implemented but aggressive trimming or filtering of sample sequences should only be used with caution of their consequences. Pseudoalignment can be a feasible choice if computational power is limited to smaller scales. The choice of tools can vary between different studies or research groups depending on accessibility to computing resources and or familiarity with certain tools or programming languages. Further studies comparing *in vivo* and *in vitro* samples using different references would be worth exploring, since there are differences between DEGs found using different genomes, and this would contribute insight to patterns in their expression and treatment of PA in clinical settings.

Acknowledgments

I would like to thank the department and Magnus Paulsson for their insight into the fantastic microbiological world of infectious disease and their unwavering encouragement and support with this project.

References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bhagirath, A. Y., Li, Y., Somayajula, D., Dadashi, M., Badr, S., & Duan, K. (2016). Cystic fibrosis lung environment and Pseudomonas aeruginosa infection. *BMC Pulmonary Medicine*, *16*(1), 1–22. https://doi.org/10.1186/S12890-016-0339-5/TABLES/2

Boucher, R. C. (2007). Evidence for airway surface dehydration as the initiating event in CF airway disease. *Journal of Internal Medicine*, *261*(1), 5–16. https://doi.org/10.1111/J.1365-2796.2006.01744.X

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology 2016 34:5*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods 2021 18:4*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Bush, S. J., Connor, T. R., Peto, T. E. A., Crook, D. W., & Walker, A. S. (2020). Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microbial Genomics*, *6*(7), 5–18. https://doi.org/10.1099/MGEN.0.000393

Cornforth, D. M., Dees, J. L., Ibberson, C. B., Huse, H. K., Mathiesen, I. H., Kirketerp-Møller, K., Wolcott, R. D., Rumbaugh, K. P., Bjarnsholt, T., & Whiteley, M. (2018). Pseudomonas aeruginosa transcriptome during human infection. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(22), E5125. https://doi.org/10.1073/PNAS.1717525115/-/DCSUPPLEMENTAL

Fan, K., Cao, Q., & Lan, L. (2021). Genome-Wide Mapping Reveals Complex Regulatory Activities of BfmR in Pseudomonas aeruginosa. *Microorganisms*, *9*(3), 1–22. https://doi.org/10.3390/MICROORGANISMS9030485

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (n.d.). *Roary: rapid large-scale prokaryote pan genome analysis*. https://doi.org/10.1093/bioinformatics/btv421

Polland, L., Rydén, H., Su, Y., Paulsson, M., (accepted: July 12, 2023). *In vivo gene expression of Haemophilus influenzae during human pneumonia.*In press.

Rossi, E., La Rosa, R., Bartell, J. A., Marvig, R. L., Haagensen, J. A. J., Sommer, L. M., Molin, S., & Johansen, H. K. (2020). Pseudomonas aeruginosa adaptation and evolution in patients with cystic fibrosis. *Nature Reviews Microbiology 2020 19:5*, *19*(5), 331–342. https://doi.org/10.1038/s41579-020-00477-5

Sarkisova, S. A., Lotlikar, S. R., Guragain, M., Kubat, R., Cloud, J., Franklin, M. J., & Patrauchan, M. A. (2014). A Pseudomonas aeruginosa EF-Hand Protein, EfhP (PA4107), Modulates Stress Responses and Virulence at High Calcium Concentration. *PLoS ONE*, *9*(6), 98985. https://doi.org/10.1371/JOURNAL.PONE.0098985

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. K. (2020). Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana. *International Journal of Molecular Sciences*, *21*(5). https://doi.org/10.3390/IJMS21051720

Scotet, V., L'hostis, C., & Férec, C. (2020). The Changing Epidemiology of Cystic Fibrosis: Incidence, Survival and Impact of the CFTR Gene Discovery. *Genes*, *11*(6). https://doi.org/10.3390/GENES11060589

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/BIOINFORMATICS/BTU153

Stephenson, A. L., Stanojevic, S., Sykes, J., & Burgel, P. R. (2017). The changing epidemiology and demography of cystic fibrosis. *La Presse Médicale*, *46*(6), e87–e95. https://doi.org/10.1016/J.LPM.2017.04.012

Tai, A. S., Bell, S. C., Kidd, T. J., Trembizki, E., Buckley, C., Ramsay, K. A., David, M., Wainwright, C. E., Grimwood, K., & Whiley, D. M. (2015). Genotypic Diversity within a Single Pseudomonas aeruginosa Strain Commonly Shared by Australian Patients with Cystic Fibrosis. *PloS One*, *10*(12). https://doi.org/10.1371/JOURNAL.PONE.0144022

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit Y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005a). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13950. https://doi.org/10.1073/PNAS.0506758102

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit Y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005b). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13950–13955. https://doi.org/10.1073/PNAS.0506758102/SUPPL_FILE/06758TABLE2.PDF

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141. https://doi.org/10.1101/gr.772403

Uluer, A., & Marty, F. M. (2014). Cystic Fibrosis. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, *1*, 874-885.e3. https://doi.org/10.1016/B978-1-4557-4801-3.00073-4

Wei, Q., & Ma, L. Z. (2013). Biofilm Matrix and Its Regulation in Pseudomonas aeruginosa. *International Journal of Molecular Sciences 2013, Vol. 14, Pages 20983-21005*, *14*(10), 20983–21005. https://doi.org/10.3390/IJMS141020983

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, *17*(1), 1–13. https://doi.org/10.1186/S12859-016-0956-2/TABLES/2

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/S13059-019-1891-0/FIGURES/2

Wu, W., Jin, Y., Bai, F., & Jin, S. (2014). Pseudomonas aeruginosa. *Molecular Medical Microbiology*, 753–767. https://doi.org/10.1016/B978-0-12-397169-2.00041-X

Supplementary Material

Supp table.1 PA strains used in the creation of the pangenome.

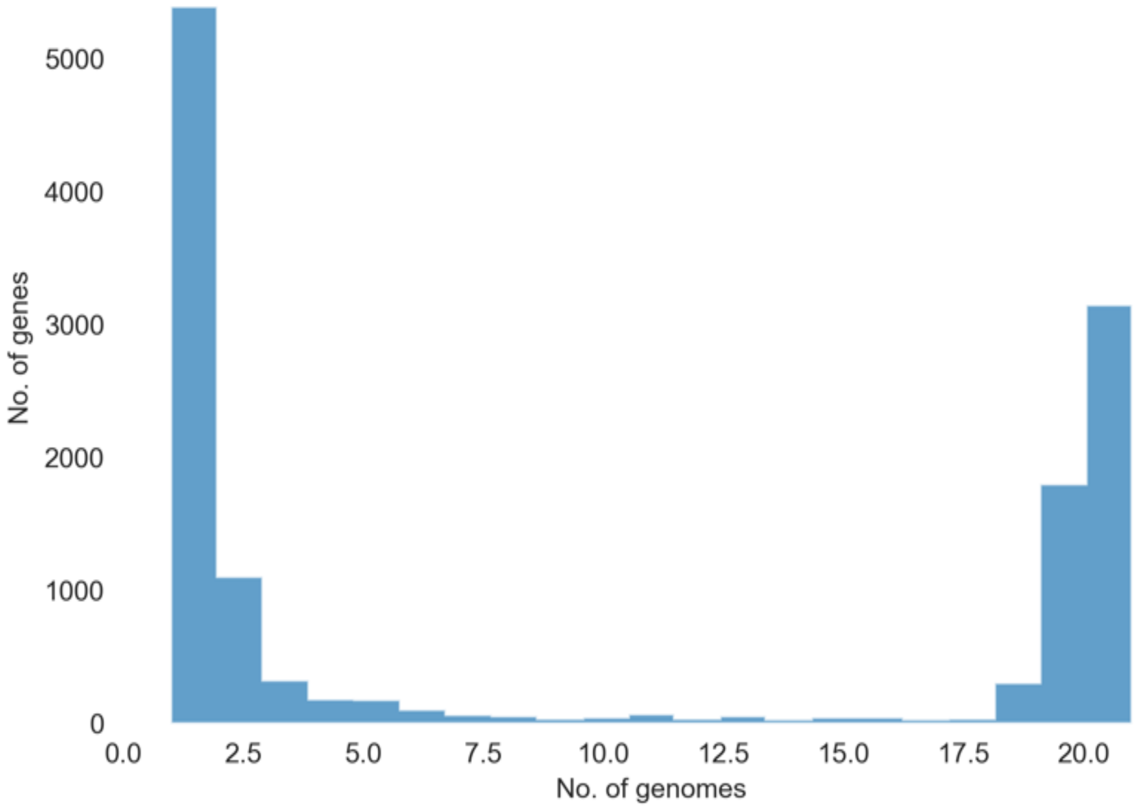
KEGG entry	Name and strain	RefSeq	GenBank	Pseudomonas.com AA file name
T00035	Pseudomonas aeruginosa PAO1	GCF_000006765.1	GCA_000006765.1	Pseudomonas_aeruginosa_PAO1_107.faa
T00401	Pseudomonas aeruginosa UCBPP-PA14	GCF_000014625.1	GCA_000014625.1	Pseudomonas_aeruginosa_UCBPP-PA14_109.faa
T00569	Pseudomonas aeruginosa PA7	GCF_000017205.1	GCA_000017205.1	Pseudomonas_aeruginosa_PA7_119.faa
T00818	Pseudomonas aeruginosa LESB58	GCF_000026645.1	GCA_000026645.1	Pseudomonas_aeruginosa_LESB58_125.faa
T01973	Pseudomonas aeruginosa M18	GCF_000226155.1	GCA_000226155.1	Pseudomonas_aeruginosa_M18_172.faa
T02161	Pseudomonas aeruginosa DK2	GCF_000271365.1	GCA_000271365.1	Pseudomonas_aeruginosa_DK2_174.faa
T01974	Pseudomonas aeruginosa NCGM2.S1	GCF_000284555.1	GCA_000284555.1	Pseudomonas_aeruginosa_NCGM2S1_173.faa
T02627	Pseudomonas aeruginosa B136-33	GCF_000359505.1	GCA_000359505.1	Pseudomonas_aeruginosa_B136-33_191.faa
T02711	Pseudomonas aeruginosa RP73	GCF_000414035.1	GCA_000414035.1	Pseudomonas_aeruginosa_RP73_192.faa
T03171	Pseudomonas aeruginosa PAO581	GCF_000468555.2	GCA_000468555.1	Pseudomonas_aeruginosa_PAO581_495.faa
T03098	Pseudomonas aeruginosa c7447m	GCF_000468935.2	GCA_000468935.1	Pseudomonas_aeruginosa_C7447m_494.faa
T03170	Pseudomonas aeruginosa PAO1-VE2	GCF_000484495.2	GCA_000484495.1	Pseudomonas_aeruginosa_PAO1-VE2_493.faa
T03097	Pseudomonas aeruginosa PAO1-VE13	GCF_000484545.2	GCA_000484545.1	Pseudomonas_aeruginosa_PAO1-VE13_492.faa
T02928	Pseudomonas aeruginosa PA1	GCF_000496605.2	GCA_000496605.2	Pseudomonas_aeruginosa_PA1_497.faa
T02929	Pseudomonas aeruginosa PA1R	GCF_000496645.1	GCA_000496645.1	Pseudomonas_aeruginosa_PA1R_496.faa

T02951	Pseudomonas aeruginosa MTB-1	GCF_000504045.1	GCA_000504045.1	Pseudomonas_aeruginosa_MTB-1_210.faa
T02970	Pseudomonas aeruginosa LES431	GCF_000508765.1	GCA_000508765.1	Pseudomonas_aeruginosa_LES431_489.faa
T02971	Pseudomonas aeruginosa SCV20265	GCF_000510305.1	GCA_000510305.1	Pseudomonas_aeruginosa_SCV20265_215.faa
T03035	Pseudomonas aeruginosa YL84	GCF_000524595.1	GCA_000524595.1	Pseudomonas_aeruginosa_YL84_2501.faa
T03031	Pseudomonas aeruginosa PA38182	GCF_000531435.1	GCA_000531435.1	Pseudomonas_aeruginosa_PA38182_7613.faa
T03789	Pseudomonas aeruginosa NCGM 1900	GCF_000829275.1	GCA_000829275.1	Pseudomonas_aeruginosa_NCGM1900_2620.faa

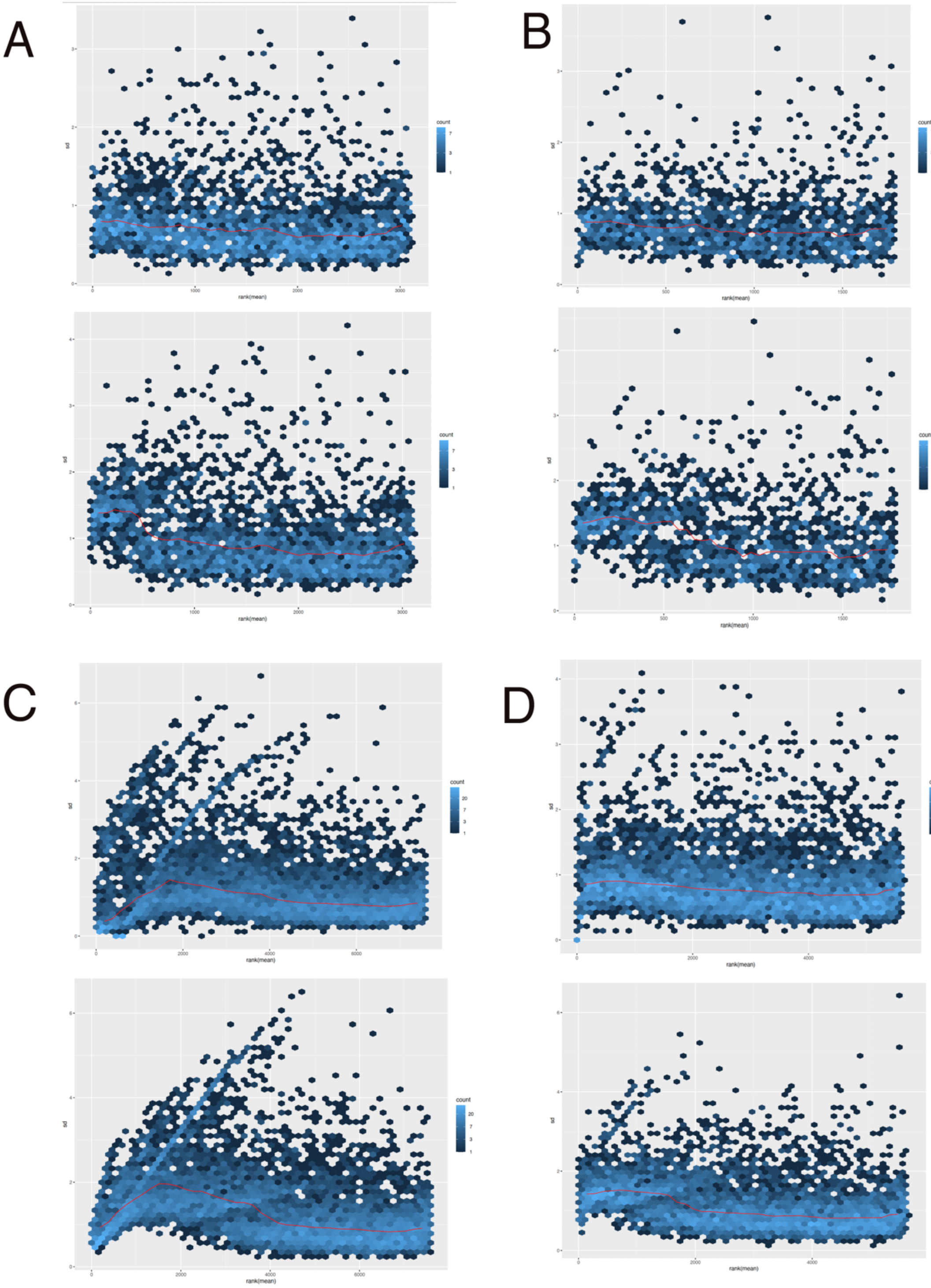
Supp table 2 Data on the sample raw reads

SRA accession	**Description**	**Sample name**
SRR6833347	Human sputum	SP01
SRR6833344	Human sputum	SP02
SRR6833345	Human sputum	SP03
SRR6833346	Human sputum	SP04
SRR6833349	Human sputum	SP05
SRR6833350	Human sputum	SP06
SRR6833351	Human sputum	SP07
SRR6833320	In vitro	INV01
SRR6833321	In vitro	INV02
SRR6833334	In vitro	INV03
SRR6833333	In vitro	INV04
SRR6833339	In vitro	INV05
SRR6833337	In vitro	INV06

Supp fig,1 Number of genes with each added genome in the pangenome



Supp fig.2. Standard deviation against mean plots for rlog (top) and vst (bottom) transformed data with core (A), soft-core (B), pangenome (C) and PAO1 (D) reference.



Supp table 3.1 Top 30 DEGs with highest LFC upregulated in *in vivo* samples (core)

Locus	**Product**	**LFC**	**padj**
PA3546	alginate biosynthesis protein AlgX	8.72735067459732	1.809181210769e-15
PA3540	GDP-mannose 6-dehydrogenase	8.72085014980398	2.41954663957224e-28

PA3557	4-amino-4-deoxy-L-arabinose-phospho-UDP flippase subunit E	7.86884925493475	2.09635079184614e-15
PA4495	hypothetical protein	7.74726583503276	1.54314164931535e-34
PA3551	mannose-1-phosphate guanylyltransferase	7.68192210917389	3.05514898658392e-25
PA4883	hypothetical protein	7.64312745859506	8.49102329151025e-48
PA3601	50S ribosomal protein L31	7.54142572063107	1.49302011578525e-123
PA3541	glycosyl transferase	7.51741477269646	1.12593717702441e-27
PA4836	hypothetical protein	7.46525470523055	2.33471862604233e-41
PA3544	alginate biosynthesis protein AlgE	7.29250654926474	3.74612471652051e-20
PA3549	alginate o-acetyltransferase AlgJ	7.2802003730797	1.23235414416003e-22
PA1318	cytochrome o ubiquinol oxidase subunit I	7.23693665274876	1.03956669065418e-20
PA3555	4-deoxy-4-formamido-L-arabinose-phospho-UDP deformylase	7.05915959932706	1.809181210769e-15
PA3284	hypothetical protein	6.97906306584287	1.25012508633195e-35
PA4837	TonB-dependent siderophore receptor family protein	6.93046874672158	9.78853986514742e-79
PA4884	hypothetical protein	6.90609029810302	3.26440349387906e-47
PA1924	hypothetical protein	6.89300156232974	2.0012452255878e-32
PA1922	TonB-dependent receptor	6.66089497707904	8.08382127754272e-51
PA3382	phosphonate ABC transporter permease	6.65498531053994	5.36133225841156e-19
PA3550	alginate o-acetyltransferase AlgF	6.62235211227699	1.21511985165431e-20
PA3556	4-amino-4-deoxy-L-arabinose transferase	6.545473863777	7.68436495597257e-25
PA3553	glycosyl transferase 2 family protein	6.52857865641282	2.08282670440422e-27
PA3887	Na+/H+ antiporter NhaP	6.45159551315514	4.19892948769198e-18
PA5536	RNA polymerase-binding protein DksA	6.3855404402549	4.15170903471729e-60
PA3547	poly(beta-D-mannuronate) lyase	6.3669538184574	8.0805474507767e-18
PA5535	hypothetical protein	6.34043728925861	8.49763328290303e-81
PA3552	UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase	6.32404544006009	5.69006569666764e-36
PA0672	heme oxygenase	6.285688453289	4.90261952307695e-39
PA2504	hypothetical protein	6.2561013688627	1.2060526918772e-29
PA3542	Mannuronan synthase	6.24411553612907	1.91617775975478e-15

Supp table 3.2 Top 30 DEGs with highest LFC upregulated in *in vivo* samples (soft-core)

Locus	Product	LFC	padj
PA2231	undecaprenyl-phosphate glucose phosphotransferase	12.0661852257856	4.08467260041555e-39
PA2232	mannose-1-phosphate guanylyltransferase	10.350806429335	4.36752212980518e-25
PA2233	glycosyl transferase	10.2146877174039	3.96061406189783e-20
PA2230	hypothetical protein	9.71690029090534	1.70782640855982e-25

PA0737	hypothetical protein	9.65577758324045	1.52342097182293e-27
PA2234	sugar ABC transporter substrate-binding protein	9.21889250014078	4.89820147987707e-18
PA1343	hypothetical protein	8.42160254109762	4.36752212980518e-25
PA4110	beta-lactamase	7.60415608498913	8.18148009086465e-63
PA2382	L-lactate dehydrogenase	7.42594116748843	1.21474306507635e-17
PA2901	hypothetical protein	7.24097604269108	3.48294069201552e-18
PA4896	RNA polymerase sigma factor	7.00944432069077	3.83966369252187e-38
PA1921	hypothetical protein	6.77297804084437	2.56035823856146e-22
PA3281	hypothetical protein	6.59760119067687	8.36141725189782e-36
PA4773	S-adenosylmethionine decarboxylase proenzyme	6.55378981356946	6.0753422456519e-15
PA2426	extracytoplasmic-function sigma-70 factor	6.54339862039224	7.75902848380523e-66
PA3283	hypothetical protein	6.53304186693394	2.27193773510739e-37
PA2137	hypothetical protein	6.46731189057167	4.59396897507776e-09
PA3558	4-amino-4-deoxy-L-arabinose-phosphoundecaprenol flippase subunit ArnF	6.44580224579697	7.04127111068815e-08
PA4471	hypothetical protein	6.34265979503045	2.63877637837442e-49
PA2412	hypothetical protein	6.20613361559376	8.28955213357481e-55
PA2114	MFS transporter	6.08296292857876	1.22370519413049e-37
PA3282	hypothetical protein	6.00675970784409	6.01092390260704e-40
PA0806	hypothetical protein	5.90028023826561	1.65169578912543e-14
PA2562	hypothetical protein	5.69252134596141	6.1521600619956e-24
PA4206	efflux transporter	5.55291263699767	1.79008951082523e-08
PA0675	RNA polymerase sigma factor	5.50224707575977	9.53821983061741e-10
PA4122	hypothetical protein	5.49384151624025	1.05307831216081e-07
PA2413	diaminobutyrate--2-oxoglutarate aminotransferase	5.49253729425439	3.3474494394307e-32
PA3237	hypothetical protein	5.42481770860905	2.00569478991252e-24
PA2468	ECF sigma factor FoxI	5.32259016136384	3.3474494394307e-32

Supp table 3.3 Top 30 DEGs with highest LFC upregulated in *in vivo* samples (pangenome)

Locus	Product	LFC	padj
PALES_27001	MerR family transcriptional regulator	29.3658884342683	3.52906826988121e-28
PADK2_24120	hypothetical protein	25.5144437993583	3.42026532465315e-16
PADK2_24115	hypothetical protein	25.2408917642357	7.17103277254137e-16
PADK2_24105	hypothetical protein	25.2124373821647	7.76117644926574e-16
PADK2_14405	hypothetical protein	24.267345251804	9.67268936046437e-15
PADK2_10875	hypothetical protein	24.1878054936276	1.19391041575833e-14
PADK2_14450	phage integrase family protein	23.9270234233675	2.38800890540655e-14
PADK2_23935	hypothetical protein	23.5890974871985	5.72659984830355e-14
PADK2_14420	outer membrane efflux protein	23.440554037947	7.95951120674698e-14
PAM18_2643	TonB-denpendent receptor	15.950825639483	2.68566574283341e-18
PADK2_08555	hypothetical protein	14.5508431064661	1.88253184530192e-11
PA4358	ferrous iron transport protein B	14.5182886912035	1.70834151895188e-35
SCV20265_1905	aminotransferase	14.2505581565393	4.4147977105467e-22
ILKJLEMH_02189	hypothetical protein	14.1696450375007	1.1346071648682e-12
PADK2_15970	hypothetical protein	13.9607091460233	1.36963872597201e-21
PALES_46081	hypothetical protein	13.7642275749314	4.4807381247146e-43
PADK2_08595	Glycosyltransferase	13.7533992601712	8.48508039970196e-21
PADK2_14190	DNA polymerase	13.6788193638487	1.15448902572632e-07
PALES_26991	ATPase P	13.5389641568962	1.30768051097447e-07
PADK2_11845	Copper-sensing two-component system response regulator CusR	13.4352397304597	1.87044039109347e-07
PADK2_10990	phage integrase	13.2807460243366	1.08549694073891e-31
PADK2_08550	UDP-N-acetyl-D-mannosaminuronate dehydrogenase	13.2614112006063	2.56718461655992e-18
PADK2_08570	hypothetical protein	13.1205150503612	5.73533372902833e-23
PA1S_RS25940	DNA-binding response regulator	13.1168250348165	4.31854991991221e-21
PSPA7_2862	cyclic peptide transporter	13.0746453235439	4.81714886460588e-18
PA4107	hypothetical protein	13.0294485476057	1.08870402771977e-18
PA4775	hypothetical protein	12.8452072138944	5.29637789118482e-21
PSPA7_4784	hypothetical protein	12.7790598333776	3.82108253107376e-25
PAM18_2607	acetyltransferase	12.7458987232445	2.30448742370938e-17
P62593	Beta-lactamase TEM	12.7261041793172	3.68725907844921e-34

Supp table 3.4 Top 30 DEGs with highest LFC upregulated in *in vivo* samples (PAO1 ref)

Locus	Product	LFC	padj
PA4107	EfhP	13.0510721578467	1.84511681741116e-15

PA4101	BfmR	12.5885038639685	8.25541655303288e-26
PA4106	conserved hypothetical protein	12.4546711175849	6.37242328847335e-12
PA2231	PslA	12.3623034402852	9.05474925716813e-42
PA4102	BfmS	12.2429604481771	4.52507527915493e-25
PA4104	conserved hypothetical protein	11.8259389904727	1.77545278611128e-13
PA3066	hypothetical protein	11.6397747498086	2.15953771097562e-32
PA0689	low-molecular-weight alkaline phosphatase B, LapB	11.5117816427607	2.75743801315848e-36
PA2220	probable transcriptional regulator	11.2847652625294	8.22857718616684e-23
PA5264	hypothetical protein	11.2485312767428	3.23684746793968e-27
PA5265	hypothetical protein	11.1703191860384	2.08657710034424e-34
PA4103	hypothetical protein	11.1333857787015	3.32163591230443e-10
PA4280.5	16S ribosomal RNA	11.1233426972329	0.000631846717268802
PA1471	hypothetical protein	11.006408809446	2.50529124908327e-24
PA2119	alcohol dehydrogenase (Zn-dependent)	10.8178956339079	1.93355213079538e-06
PA2232	PslB	10.6889133003309	2.75286921382355e-26
PA0100	hypothetical protein	10.6302615655605	1.65314275926554e-11
PA0498	hypothetical protein	10.621997847585	3.098864012077e-31
PA2233	PslC	10.6131772956741	7.07951192024871e-21
PA0497	hypothetical protein	10.3253697662804	6.85658606697171e-26
PA2456	hypothetical protein	10.3203107253949	1.84762790209762e-09
PA2771	diguanylate cyclase with a self-blocked I-site, Dcsbis	10.1676775779985	6.6346185760475e-29
PA0257	hypothetical protein	10.0955242130532	4.97563254032472e-28
PA3065	hypothetical protein	10.0867328714128	4.18517235428844e-23
PA2230	hypothetical protein	10.0662948986042	2.2525576145905e-27
PA4105	hypothetical protein	10.0374397971351	3.34156281218806e-10
PA3067	probable transcriptional regulator	9.98113228054582	3.47226998020911e-22
PA0737	hypothetical protein	9.86960275608147	2.9783305999352e-30
PA4195	putative amino acid ABC transporter substrate-binding protein	9.84591153196641	1.1579845470094e-25
PA2772	hypothetical protein	9.83938002348575	4.24397335618275e-18

Supp table 4.1 GO enrichment of upregulated DEGS in *in vivo* samples (core)

GO biological process complete	**PA - REFLIST (5564)**	**Count (882)**	**Expected**	**Over/Under represented (+/-)**	**Fold Enrichment**	**Raw P-value**	**FDR**
alginic acid metabolic process (GO:0042120)	16	13	2.54	+	5.13	4.39E-05	9.76E-02
monoatomic ion transport (GO:0006811)	71	29	11.25	+	2.58	4.95E-05	5.51E-02
monoatomic cation transport (GO:0006812)	67	28	10.62	+	2.64	5.48E-05	4.07E-02

Supp table 4.2 GO enrichment of upregulated DEGS in *in vivo* samples (softcore)

GO biological process complete	PA - REFLIST (5564)	Count (882)	Expected	Over/Under represented (+/-)	Fold Enrichment	Raw P-value	FDR
Unclassified (UNCLASSIFIED)	3254	330	278.96	+	1.18	4.67E-06	3.46E-03
biological_process (GO:0008150)	2310	147	198.04	-	.74	4.67E-06	2.60E-03
metabolic process (GO:0008152)	1293	77	110.85	-	.69	3.20E-04	3.23E-02
organic substance metabolic process (GO:0071704)	1217	71	104.33	-	.68	2.37E-04	3.11E-02
cellular process (GO:0009987)	1657	95	142.05	-	.67	3.42E-06	7.60E-03
nitrogen compound metabolic process (GO:0006807)	908	49	77.84	-	.63	3.18E-04	3.37E-02
primary metabolic process (GO:0044238)	971	48	83.24	-	.58	1.79E-05	7.96E-03
organic substance biosynthetic process (GO:1901576)	672	31	57.61	-	.54	1.37E-04	2.77E-02
biosynthetic process (GO:0009058)	681	31	58.38	-	.53	8.25E-05	2.04E-02
cellular biosynthetic process (GO:0044249)	573	25	49.12	-	.51	2.17E-04	3.22E-02
organonitrogen compound biosynthetic process (GO:1901566)	407	15	34.89	-	.43	2.40E-04	2.96E-02

macromolecule metabolic process (GO:0043170)	518	17	44.41	-	.38	4.60E-06	5.12E-03
nitrogen compound transport (GO:0071705)	256	7	21.95	-	.32	4.12E-04	3.99E-02
small molecule biosynthetic process (GO:0044283)	232	4	19.89	-	.20	3.87E-05	1.43E-02
gene expression (GO:0010467)	158	2	13.55	-	.15	2.97E-04	3.31E-02
protein transport (GO:0015031)	130	1	11.14	-	.09	4.31E-04	4.00E-02
establishment of protein localization (GO:0045184)	135	1	11.57	-	.09	2.92E-04	3.42E-02
cellular macromolecule localization (GO:0070727)	142	1	12.17	-	.08	2.06E-04	3.52E-02
protein localization (GO:0008104)	142	1	12.17	-	.08	2.06E-04	3.27E-02
cellular localization (GO:0051641)	156	1	13.37	-	.07	6.15E-05	1.71E-02
cellular component organization or biogenesis (GO:0071840)	161	1	13.80	-	.07	4.14E-05	1.31E-02
amino acid biosynthetic process (GO:0008652)	111	0	9.52	-	< 0.01	2.27E-04	3.15E-02
protein transmembrane transport (GO:0071806)	112	0	9.60	-	< 0.01	1.40E-04	2.60E-02
alpha-amino acid biosynthetic process (GO:1901607)	99	0	8.49	-	< 0.01	4.89E-04	4.36E-02

cellular component biogenesis (GO:0044085)	119	0	10.20	-	< 0.01	9.39E-05	2.09E-02

Supp table 4.3 GO enrichment of upregulated DEGS in *in vivo* samples (pan)

GO biological process complete	PA - REFLIST (5564)	Count (882)	Expected	Over/Under represented (+/-)	Fold Enrichment	Raw P-value	FDR
cellular nitrogen compound metabolic process (GO:0034641)	557	93	157.57	-	.59	2.23E-07	4.95E-04
nucleobase-containing compound metabolic process (GO:0006139)	327	44	92.51	-	.48	3.10E-07	3.45E-04
nitrogen compound metabolic process (GO:0006807)	908	178	256.86	-	.69	6.39E-07	4.74E-04
organonitrogen compound biosynthetic process (GO:1901566)	407	64	115.14	-	.56	2.15E-06	1.20E-03
translation (GO:0006412)	80	3	22.63	-	.13	3.47E-06	1.54E-03
nucleic acid metabolic process (GO:0090304)	207	24	58.56	-	.41	3.57E-06	1.32E-03
macromolecule metabolic process (GO:0043170)	518	91	146.54	-	.62	5.08E-06	1.61E-03
heterocycle metabolic process (GO:0046483)	463	79	130.98	-	.60	7.12E-06	1.98E-03
primary metabolic process (GO:0044238)	971	203	274.69	-	.74	1.31E-05	3.25E-03
biosynthetic process (GO:0009058)	681	132	192.65	-	.69	1.54E-05	3.42E-03
organic substance biosynthetic process (GO:1901576)	672	131	190.10	-	.69	2.10E-05	4.24E-03
gene expression (GO:0010467)	158	17	44.70	-	.38	2.66E-05	4.92E-03
carbohydrate derivative metabolic process (GO:1901135)	202	26	57.14	-	.45	3.13E-05	5.36E-03

organic cyclic compound metabolic process (GO:1901360)	501	95	141.73	-	.67	1.29E-04	2.06E-02
cellular biosynthetic process (GO:0044249)	573	113	162.10	-	.70	1.55E-04	2.30E-02
cellular component organization or biogenesis (GO:0071840)	161	20	45.55	-	.44	1.73E-04	2.41E-02
cellular nitrogen compound biosynthetic process (GO:0044271)	363	64	102.69	-	.62	1.84E-04	2.41E-02
transition metal ion transport (GO:0000041)	36	27	10.18	+	2.65	1.94E-04	2.39E-02
ncRNA metabolic process (GO:0034660)	75	5	21.22	-	.24	1.98E-04	2.32E-02
iron ion transport (GO:0006826)	30	24	8.49	+	2.83	2.18E-04	2.42E-02
O antigen metabolic process (GO:0046402)	35	0	9.90	-	< 0.01	2.87E-04	3.04E-02
O antigen biosynthetic process (GO:0009243)	35	0	9.90	-	< 0.01	2.87E-04	2.90E-02
cellular component biogenesis (GO:0044085)	119	13	33.66	-	.39	2.88E-04	2.78E-02
cellular aromatic compound metabolic process (GO:0006725)	469	90	132.68	-	.68	2.97E-04	2.75E-02
iron coordination entity transport (GO:1901678)	23	20	6.51	+	3.07	3.36E-04	2.99E-02
organonitrogen compound metabolic process (GO:1901564)	682	143	192.93	-	.74	4.15E-04	3.55E-02

Supp table 4.4 GO enrichment of upregulated DEGS in *in vivo* samples (PAO1 ref)

GO biological process complete	**PA - REFLIST (5564)**	**Count (882)**	**Expected**	**Over/ Under represented (+/-)**	**Fold Enrichment**	**Raw P-value**	**FDR**
cellular nitrogen compound metabolic process (GO:0034641)	557	93	161.67	-	.58	5.30E-08	1.18E-04

nucleobase-containing compound metabolic process (GO:0006139)	327	43	94.91	-	.45	6.13E-08	6.82E-05
nucleic acid metabolic process (GO:0090304)	207	22	60.08	-	.37	3.10E-07	2.30E-04
nitrogen compound metabolic process (GO:0006807)	908	187	263.55	-	.71	1.90E-06	1.06E-03
gene expression (GO:0010467)	158	15	45.86	-	.33	1.93E-06	8.58E-04
translation (GO:0006412)	80	3	23.22	-	.13	2.16E-06	8.03E-04
organonitrogen compound biosynthetic process (GO:1901566)	407	68	118.14	-	.58	4.92E-06	1.56E-03
heterocycle metabolic process (GO:0046483)	463	81	134.39	-	.60	5.22E-06	1.45E-03
cellular component organization or biogenesis (GO:0071840)	161	17	46.73	-	.36	6.91E-06	1.71E-03
cellular component biogenesis (GO:0044085)	119	10	34.54	-	.29	1.01E-05	2.26E-03
macromolecule metabolic process (GO:0043170)	518	96	150.35	-	.64	1.31E-05	2.64E-03
biosynthetic process (GO:0009058)	681	136	197.67	-	.69	1.52E-05	2.81E-03
organic substance biosynthetic process (GO:1901576)	672	135	195.05	-	.69	2.07E-05	3.54E-03
primary metabolic process (GO:0044238)	971	214	281.84	-	.76	5.33E-05	8.48E-03
carbohydrate derivative metabolic process (GO:1901135)	202	28	58.63	-	.48	7.47E-05	1.11E-02
cellular nitrogen compound biosynthetic process (GO:0044271)	363	64	105.36	-	.61	7.68E-05	1.07E-02
cellular biosynthetic process (GO:0044249)	573	115	166.32	-	.69	9.79E-05	1.28E-02
organic cyclic compound metabolic process (GO:1901360)	501	98	145.42	-	.67	1.23E-04	1.52E-02
ncRNA metabolic process (GO:0034660)	75	5	21.77	-	.23	1.30E-04	1.53E-02
transition metal ion transport (GO:0000041)	36	28	10.45	+	2.68	1.34E-04	1.50E-02

iron ion transport (GO:0006826)	30	25	8.71	+	2.87	1.44E-04	1.52E-02
RNA processing (GO:0006396)	55	2	15.96	-	.13	1.61E-04	1.62E-02
iron coordination entity transport (GO:1901678)	23	21	6.68	+	3.15	2.05E-04	1.98E-02
cellular aromatic compound metabolic process (GO:0006725)	469	92	136.13	-	.68	2.24E-04	2.07E-02
ncRNA processing (GO:0034470)	53	2	15.38	-	.13	2.42E-04	2.15E-02
O antigen metabolic process (GO:0046402)	35	0	10.16	-	< 0.01	2.97E-04	2.54E-02
O antigen biosynthetic process (GO:0009243)	35	0	10.16	-	< 0.01	2.97E-04	2.45E-02
cellular component organization (GO:0016043)	121	14	35.12	-	.40	3.56E-04	2.83E-02
RNA metabolic process (GO:0016070)	144	19	41.80	-	.45	4.24E-04	3.25E-02
organelle organization (GO:0006996)	40	1	11.61	-	.09	5.51E-04	4.09E-02
cellular process (GO:0009987)	1657	411	480.96	-	.85	6.69E-04	4.80E-02