Resistance development of *Pseudomonas aeruginosa* over the past half-century

Report Theme 3

Course I0I31a: Integrated Practicals of Molecular Biotechnology (IPMB)

Group C

Kloeck Arend

Roesgen Roxane

Vancompernolle Robynn

Verdonck Marie

Studying resistance development in Pseudomonas aeruginosa

Introduction

In the last century, antibiotic resistance has been a major challenge in global health. Much research has been performed to identify the major causes of global antibiotic resistance development. Additionally, alternative approaches such as vaccines or antibodies are increasingly explored. From a socioeconomic and political point of view, an interesting distinction can be made between developing and developed countries (Chokshi *et al*., 2019). In developing countries, clinical misuse, low quality and high availability of antibiotics, and lack of surveillance of resistance development have been identified as contributing factors. In developed countries, the major factors are the excessive use of antibiotics in food-producing animals and poor hospital-based regulations. In essence, the misuse and excessive use of antibiotics are what lead to resistance development. Antibiotics are frequently prescribed needlessly, for instance, to treat diseases like the flu and the common cold that don't respond to antibiotics. Furthermore, overuse of antibiotics creates a selective pressure that favors the development of resistant bacteria, thus creating stronger germs (DHSS, n.d.). One significant species for which antibiotic resistance has been highlighted as a major healthcare issue is *Pseudomonas aeruginosa* (Langendonk *et al.*, 2021).

In this work, resistance development of *P. aeruginosa* to aminoglycosides and fluoroquinolones, two classes of antibiotics, is examined by whole-genome analysis. The use of Illumina HiSeq2500, a next-generation sequencing platform widely used in genomic research, and paired-end sequencing data is explored. Three strains isolated before 1940 are compared to three other strains isolated after 1940 for the presence of specific mutations and proteins. For this purpose, three draft genomes of the sequenced strains isolated after 1940 are constructed and compared to three NCBI assemblies of the strains isolated before 1940. For aminoglycoside resistance, the presence of aminoglycoside-modifying enzymes (AMEs) is examined. For fluoroquinolone resistance, the DNA gyrase subunit A (GyrA) and DNA topoisomerase subunit A (ParC) are studied.

AMEs are antibiotic-modifying enzymes whose presence is often linked to aminoglycoside resistance. By catalyzing modifications at different hydroxyl and amino groups, these enzymes inactivate the antibiotics (Egorov *et al*., 2018). In strains lacking these AMEs, increased cell wall impermeability has been identified as a resistance mechanism (Barnhart et al., 2002). GyrA and ParC form a catalytic site involved in breaking DNA molecules. Fluoroquinolone resistance has been linked to reduced efficiency of the interaction of these molecules with the DNA gyrase - DNA topoisomerase complex (Egorov *et al*., 2018). The reduction in interaction efficiency is often due to mutations in GyrA and/or ParC, altering the fluoroquinolones binding pocket of the complex.

The most interesting findings of this study are that one AME protein, an aminoglycoside 3'-phosphotransferase, was present in two of the three old strains and remained unchanged in all the new strains. As opposed to looking at fluoroquinolone resistance, where this study has found a large difference between the new and old strains.

Methods

Six different strains were used (Table A.I.1 and Table A.I.1). Raw sequencing data generated by Illumina HiSeq 2500 was retrieved from the European NGS sequence database ENA for the strains **Aa249, co380791, 1709-12**. Both forward (R1) and reverse (R2) reads were downloaded to acquire paired-end data, which reveals more information about each DNA fragment than reads sequenced solely by single-end sequencing. For each read, a sequence identifier, nucleotide sequence, separator, and quality scores based on the Phred scoring system ASCII_Base +33 are provided.

Different sets of analyses were exploited to assess the quality of the raw sequencing data using the software Galaxy. The **per base sequence quality scores** represent the quality of the reads for every single base. The higher the score, the more accurate the base call. The **per sequence quality scores** plot displays the average quality score per read for all the reads in the analysis. High-quality data exhibits a tight peak at a high mean sequence quality score and no peaks at lower values. **The per base sequence content** plots the percentage of nucleotides at each position in the read. High-quality data should display an approximately constant percentage of each base over the entire read, with

little or no difference between the occurrence of all four bases. The **per sequence GC content** represents the proportionality of G and C per read. A uniform GC content for all reads is expected in high-quality data (Babraham Bioinformatics, n.d.). The **per base N content** represents the proportion of positions in the read for which no base was called. For high-quality data, this value should be close to or equal to zero. The **sequence length distribution** plot reveals the distribution of sequence lengths over all sequences in the analysis. The plot often shows a single peak, at a specific length. The **sequence duplication levels** reveal how often every sequence occurs in the analysis. In high-quality data sets, most sequences occur once. **Overrepresented sequences** should not appear in libraries containing diverse sequences. The **adapter content** shows the percentage of reads with different adapters per position in the read. High-quality data should not have any adapter sequences. See Appendix II for a complete set of information on the parameters used for quality assessment.

To ensure downstream analysis would not be influenced by low-quality reads, the reads were subjected to different 'trim and clean up' operations in Trimmomatic. To improve the adapter content of the reads, ILLUMINACLIP was performed emphasizing the TruSeq3 (paired-ended, for MiSeq & HiSeq) adapter, consistent with the HiSeq 2500 data used in this experiment. A detailed overview of the parameters used is provided in Appendix III. Subsequently, a headcrop operation was performed removing five bases from the start of the read to improve the per base sequence content that deviated most strongly at these positions. For each strain, four output files are generated. A paired forward and reverse fastq file, and an unpaired forward and reverse fastq file. Both reads were kept or dropped if they did not meet the criteria after trimming. Forward or reverse only surviving reads were removed to work with paired-end data solely.

A *de novo* assembly was performed in SPAdes using the trimmed paired reads as inputs ('only assembler'). The k-mer size was automatically chosen by SPAdes which assembles k-mers of different values (21, 33, 55) to select the assembly with the best N50 score (explained in paragraph hereunder (Page *et al*., 2016). K-mers are short DNA fragments of length k, used to assemble a larger genome based on the overlap of shorter (unique) fragments (Skliros, 2018). SPAdes provides a file with contigs, continuous sequences assembled from a larger set of sequence fragments, and a file with scaffolds, which consist of contigs joined in a specific relative position/orientation, erratically separated by gaps (PacBio, 2023).

The quality of the assembly was assessed in Quast, which provides assembly metrics. The **number of contigs** with a length greater than 0 and 1000 bp are given, as well as the **total length** which represents the total number of bases in contigs of a length greater or equal to 0 and 1000 (CAB, n.d.). The **number of contigs** is the total number of contigs in the final assembly, and the length of the **largest contig** is also provided. The **N50/90** is the length of the smallest contigs at 50/90% of the total length. The **L50/90** value represents the smallest number of contigs needed to include at least 50/90% of all nucleotides in the assembly (NCBI, n.d.a). A visual representation of the N50 and L50 values is given in Figure A.IV.1. The **number of N's per 100 kb** represents the average number of uncalled bases. The generated assemblies are compared to the reference genomes of *P. aeruginosa* strains PA14 (GCA_025490475.1) and POA1 (GCA_003957105.1).

To annotate the assemblies, the software Prokka was used. For strains **A15**, **A17**, and **A22**, the assemblies were retrieved from NCBI. For strains **Aa249**, **co380791**, and **1709-12**, the assemblies generated in the previous steps of this experiment were used. These protein annotations were used to build a protein database in NCBI Blast+. To search for proteins related to antibiotic resistance, a BLAST search against protein sequences of known antibiotic resistance-associated proteins was performed. For the aminoglycoside resistance, 149 protein sequences of AMEs from different bacteria were used as query sequences, and the generated protein database as the subject database. When searching a database of a particular size, one can expect to obtain a number of hits purely by chance. This is reflected by the Expect (E) value (NCBI, n.d.b). To obtain matches of very high quality, an E-value of 10^{-145} was used in a traditional BLASTP. For the fluoroquinolone resistance, the protein sequences of GyrA (NP_251858.1, 923 aa) and ParC (NP_253651.1, 754 aa) of the reference genome *P. aeruginosa* (NC_002516.2) were used as query. To obtain only one significant match per strain, a stringent E-value of 10^{-128} was used for GyrA and for ParC in a traditional BLASTP. A multiple sequence alignment is performed on the protein sequences of ParC and GyrA in MAFFT and subsequently used

to find SNP sites, allowing analysis of possible amino acid substitutions associated with fluoroquinolone resistance.

Results

Analysis and quality assessment of raw sequencing data

Raw sequencing data was obtained for Aa249, co380791, and 1709-12. An overview of the data can be found in Table 1. The top half of the table shows the total amount of sequences, sequence length, and GC content. The bottom half shows the Phred quality score of the 5th nucleotide of the first R1 read of each strain, as well as the accuracy of this nucleotide.

Table 1: Overview raw sequencing data Aa249, co380791, and 1709-12.

	Aa249	**co380791**	**1709-12**
Total Sequences	507 734	1 571 394	1 790 396
Sequence Length (bp)	26-150	26-150	26-150
%GC	65	61	61
Phred Symbol/Score	#/35	D/68	D/68
Accuracy (%)	36.904	99.968	99.968

Quality assessment analysis in Galaxy, with both the R1 and R2 of the strain Aa249 as input, ascertained adequate per base sequence quality, per sequence quality scores, per base sequence content, per base N content, sequence duplication levels, and overrepresented sequences. The per sequence GC content deviates from the theoretical distribution, displaying a higher and sharper peak at the center of this distribution. The sum of the deviations from the normal distribution represents 15-30% of the reads. The sequence length distribution indicates reads of different lengths in the data set (Range 26-150 bp) (blast2go, n.d.). The Illumina Universal Adapter (IUA) content increases up to roughly 10% and 8% for the R1 and R2, respectively. For both reads, the adapter content increases from positions 96/97 to 134/135. The three last-mentioned parameters are targets for quality improvement.

Quality assessment of co380791 ascertained adequate per base sequence quality, per sequence quality scores, per base N content, sequence duplication levels, and overrepresented sequences. The per base sequence content exhibits a deviation at positions 1-9 while improving vastly for the remainder of the reads. The difference between A and T, as well as G and C, reaches a value greater than 10% within this range of positions. The per sequence GC content differs from the theoretical distribution, displaying a higher and slightly sharper peak at the center of this distribution. The sum of the deviations from the normal distribution represents 15-30% of the reads. The sequence length distribution indicates reads of different lengths (Range 26-150 bp). The IUA content increases up to roughly 22% and 20% for the R1 and R2, respectively, from positions 36/37 to 134/135. The four last-mentioned parameters are targets for quality improvement, emphasizing the adapter content.

Quality assessment of 1709-12 ascertained adequate per base sequence quality, per sequence quality scores, per base N content, sequence duplication levels, and overrepresented sequences. In the R1, the per base sequence content deviates slightly at positions 1-9 and stabilizes afterward. In the R2, similar deviations at positions 1-9 are observed. However, the difference between G and C reaches a value greater than 10% within this range of positions. Additionally, a stronger deviation is observed in the latter part of the reads. In positions 49-150, the G content differs from the C content, and the four lines are not displaying a parallel profile (Figure 2a). The per sequence GC content deviates from the theoretical distribution, displaying a higher and slightly sharper peak at the center of this distribution. The sum of the deviations from the normal distribution represents 15-30% of the reads. The sequence length distribution indicates reads of different lengths (Range 26-150 bp). The IUA content increases up to roughly 23% and 20% for the R1 and R2, respectively, from positions 36/37 to 134/135. The three last-mentioned parameters are targets for quality improvement of the

sequencing data, emphasizing the adapter content. Additionally, the per base sequence content of the R2 is subject to further analysis and improvement.

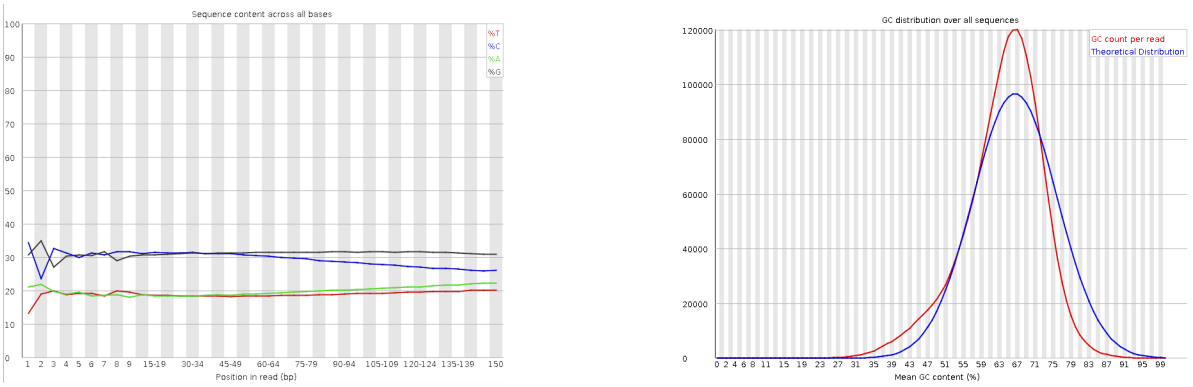

Figure 2: a) Sequence content across all bases for R1 of 1709-1. b) Per sequence GC distribution of R1 of 1709-1.

Improve read quality and clip adapters

Trimmed and processed sequencing data was obtained for Aa249, co380791, and 1709-12. Table X provides an overview of the total amount of sequences, sequence length, and GC content. Data from this table can be compared with the top half of Table 2 which shows an overview of the unprocessed reads.

Table 2: Overview of processed sequencing data Aa249, co380791, and 1709-12.

	Aa249	**co380791**	**1709-12**
Total Sequences	507 730	1 571 044	1 790 020
Sequence Length (bp)	1-145	1-145	1-145
%GC (F/R)	65/65	62/61	62/62

Four pairs of reads were removed from the Aa249 dataset, representing less than 0.001%. For the remaining 99.999%, both reads were kept. Quality assessment on the filtered data ascertained adequate per base sequence quality, per sequence quality scores, per base sequence content, per base N content, sequence duplication levels, and overrepresented sequences. The per sequence GC content still deviates from the theoretical distribution, displaying a higher and sharper peak at the center of this distribution, which remained at 65%. The sum of the deviations from the normal distribution represents 15-30% of the reads. The sequence length distribution indicates reads of different lengths (Range 1-145 bp). The adapter content was reduced to 0%.

350 pairs of reads were removed from the co380791 dataset, representing 0.02%. For the remaining 99.98%, both reads were kept. Quality assessment ascertained adequate per base sequence quality, per sequence quality scores, per base N content, sequence duplication levels, and overrepresented sequences. The per base sequence content was strongly reduced. The difference between A and T, as well as G and C, does not reach a value greater than 10%. The per sequence GC content still differs from the theoretical distribution, displaying a higher and slightly sharper peak at the center of this distribution, which remained at 61% for the R1 but shifted to 62% for the R2. The sum of the deviations from the normal distribution represents 15-30% of the reads. The sequence length distribution indicates reads of different lengths (Range 1-145 bp). The adapter content was reduced to 0%.

376 pairs of reads were removed from the 1709-12 dataset, representing 0.02%. For the remaining 99.98%, both reads were kept. Quality assessment ascertained adequate per base sequence quality, per sequence quality scores, per base N content, sequence duplication levels, and overrepresented sequences. The per base sequence content was improved. The difference between G and C remains

lower than 10% for the R1 and R2s. The deviation previously observed in the latter part of the reads in the R2 set was removed. The per sequence GC content still deviates from the theoretical distribution, displaying a higher and slightly sharper peak at the center of this distribution, now increased to 66% (Figure 2b). The sum of the deviations from the normal distribution represents 15-30% of the reads. The sequence length distribution indicates reads of different lengths (Range 1-145 bp). The adapter content was reduced to 0%.

De novo assembly & Quality assessment

For strain Aa249, 250 contigs were assembled into 212 larger scaffolds. The length of the longest contig is 404 353 bp. The coverage of this contig is 10.137745. For strain co380791, 523 contigs were assembled into 480 larger scaffolds. The length of the longest contig is 373 509 bp. The coverage of this contig is 20.139830. For 1709-12, 448 contigs were assembled into 419 larger scaffolds. The length of the longest contig is 311 029 bp. The coverage of this contig is 25.616376.

For each strain, the contig and scaffold file differ in various statistics. The total length, N50/90, and the auN value are greater for the scaffolds whereas the L50/90 is lower. The number of mismatches equals 0 for the contigs but increases up to 1100 (Aa249 & 1709-12) and 1320 (co380791) for the scaffolds. A detailed table is provided for strains Aa249, co380791, and 1709-12 in Appendix V. The length of the draft genomes generated in SPAdes, as well as the GC content, are displayed and compared to the reference genomes of *P. aeruginosa* in Table 3. The GC contents of the generated assemblies are smaller than the reference genome. The length of both the contigs and scaffolds assemblies is larger than the reference.

Table 3: A comparison of reference genomes PA14, PAO1, and the assemblies generated for the new strains.

	Genome Length (C/S) (bp)	**GC Content (%)**	**N50 (C/S)**	**L50 (C/S)**
Ref. Genome PA14	6 535 621/6 537 581	66.30	1 640 262/6 537 581	2/1
Ref. Genome PAO1	6 218 048/6 218 048	66.60	277 905/6 218 048	8/1
Aa249	7 080 024/7 081 048	65.85	115 197/158 361	19/13
co380791	7 228 486/7 229 660	65.66	104 008/177 679	21/13
1709-12	7 050 650/7 051 764	65.83	118 419/154 612	22/15

Table 3 reveals that the reference genomes and generated assemblies differ not only in the assembly statistics N50 and L50 but also in genome length and GC content. In contrast to the reference genomes, the assemblies have rather high L50 contig and scaffold values and low N50 values. The GC content differs a little between the generated assemblies. These values are all slightly lower than the reference genomes. The lengths of the assemblies are greater than the reference genomes. The assemblies have similar lengths, N50, and L50 values. co380791 has a somewhat longer genome length and a smaller N50 value.

Search protein database for proteins related to antibiotic resistance

A custom protein database of 37 043 sequences was created. Table 4 shows the results obtained in the blast search of the AME file against the database. 19 matches were acquired. For the strain 1709-12, the enzyme Streptomycin 3'-adenylyltransferase had 14 matches (Different amounts of mismatches and E-values). The other strains, except for A22, had a match for Aminoglycoside 3'-phosphotransferase (APH). Detailed results of the analysis are set out in Table A.VI.1.

Table 4: Part of the output of the blastp search of the AME file against the generated protein database. Columns 1: Query Seq-id, 2: Subject Seq-id, 3: Percentage of identical matches, 4: Alignment length, 5: Number of mismatches, 6: E-value. 7: Protein.

ag:CAM98048	1709-12_05732	87.645	259	32	1.63e-167	1709-12_05732 Streptomycin 3''-adenylyltransferase
ag:CAA62365	1709-12_06126	98.507	268	4	0.0	1709-12_06126 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	co380791_04948	98.507	268	4	0.0	co380791_04948 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	Aa249_02951	98.507	268	4	0.0	Aa249_02951 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	A17_05014	98.507	268	4	0.0	A17_05014 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	A15_01328	97.388	268	7	0.0	A15_01328 Aminoglycoside 3'-phosphotransferase

In the blast search for ParC and GyrA, four and three protein sequences were obtained with at least one mismatch. Detailed results of the analyses are set out in Tables A.VI.2 and A.VI.3. A multiple sequencing alignment on ParC and GyrA was performed (Figures A.VI.1 & A.VI.2). The results obtained from the SNP analysis on both proteins are presented in Table 5 and 6. Three different SNPs were identified for ParC. Two SNPs are present in the old strains (A15, A22), and one in the new strains (1709-12, co380791). Six different SNPs were identified for GyrA, only present in the new strains.

Table 5: SNP analysis on ParC. Columns 1: Position of SNP. 2-3: Reference and altered amino acid. 4-9: Strains

POS	REF	ALT	FORMAT	A22	A17	A15	Aa249	1709-12	co380791
87	S	L	GT	0	0	0	0	1	1
331	S	T	GT	0	0	1	0	0	0
646	V	L	GT	1	0	0	0	0	0

Table 6: SNP analysis on GyrA. Columns 1: Position of SNP. 2-3: Reference and altered amino acid. 4-9: Strains

POS	REF	ALT	FORMAT	A22	A17	A15	Aa249	1709-12	co380791
83	T	I	GT	0	0	0	1	1	1
671	V	I	GT	0	0	0	0	1	1
860	G	S	GT	0	0	0	0	1	1
893	D	E	GT	0	0	0	0	1	1
900	A	G	GT	0	0	0	0	1	1
903	S	A	GT	0	0	0	0	1	1

Discussion

Sequencing technologies have revolutionized the field of genomics by providing researchers with high-throughput sequencing data in a cost-effective manner. One such platform that has greatly increased in popularity as a result of its high accuracy and speed is Illumina HiSeq2500. With a read length of up to 2 x 250 bps, this next-generation sequencing device can produce up to 1 billion reads each run using a sequencing-by-synthesis method. HiSeq2500 can process up to 1 TB of data each run with an accuracy rate of 99.9% (Illumina, 2023). It can be used in single- or paired-end mode. Applications include metagenomics, whole-genome, exome, and transcriptome sequencing. Other sequencing platforms include PacBio Sequel II (3rd-generation), Oxford Nanopore MinION (3rd-generation), and Ion Torrent Ion S5 (Next-generation). PacBio Sequel II employs a single-molecule real-time strategy. It can read up to 100 kb and has a high accuracy rate of 99.9% (DNA Technologies Core, n.d.). The platform, nonetheless, has a high cost per base and a low throughput. Oxford Nanopore MinION uses a nanopore sequencing technology. It is capable of producing long reads in real time with a length of up to 2 Mb. Due to high error rates of up to 15%, reads must be rectified using computational techniques (Oxford Nanopore Technologies, n.d.). Ion Torrent Ion S5 employs semiconductor sequencing technology. It produces up to 10 million reads each run with a read length of up to 600 bps (ThermoFisher, n.d.). It is inadequate for *de novo* assembly due to its high error rate. In closing, Illumina HiSeq2500 continues to be a popular choice because of its low cost and ability to produce high-quality data for a range of applications.

The process of paired-end sequencing entails sequencing both ends of a DNA fragment, producing two sequences that are separated by a defined distance. The data from paired-end sequencing is crucial for de novo genome assembly as a more accurate and complete genome assembly becomes possible by sequencing both ends of DNA fragments. Because the additional read can correct mistakes in the first read, paired-end sequencing data has higher accuracy than single-end data. Paired-end sequencing does, though, have several drawbacks, including a higher price and a longer sequencing time. Paired-end sequencing data also needs additional processing power and computational time to process and analyze (Illumina, n.d.).

A lot of data is now publicly available due to recent advancements in sequencing. The quality of this data however has to be checked and actions have to be taken accordingly. Therefore, raw sequencing data were subjected to different 'trim and clean up' operations in Trimmomatic. In general, the quality of the raw sequences was decent. The paired-end reads of Aa249, co380791, and the R1 of 1709-12 encountered warnings about 'per base sequence content', 'per base sequence GC content', 'sequence length distribution', and adaptor content. Warnings are not necessarily problematic, but when FASTQC warns for unusual data this should be corrected. I.e. the adaptors had to be clipped because more than 10% of reads contained the same sequence (blast2go, n.d.). Using the Trimmomatic operations, the quality was ameliorated in such a way that only the 'per base sequence GC content' still gave a warning. A slight shift of the curve can be attributed to a systemic bias, whereas a usual shape indicates a contaminated library. In the data used in this experiment, the warning is not problematic as there is only one peak (Figure 2b). Thus, the single peak represents a single average GC content. The curve is only slightly shifted, it also got smaller, displaying a sharper peak (blast2go. n.d.). This was the same looking distribution for all of our data, only one figure is inserted here. The data, therefore, was adequate for further usage.

Genome assembly, which enables the reconstruction of a complete genome sequence, is a crucial step in the field of genomics. The two main strategies used for genome assembly are *de novo* assembly and mapping to a reference genome. *De novo* assembly entails completely rebuilding the genome sequence from start, with no knowledge of the genome sequence beforehand. This procedure is usually carried out when a reference genome is either unavailable, of poor quality, or inaccurately represents the genome of interest (Illumina, 2015). Mapping to a reference genome involves aligning sequencing data to a pre-existing reference genome. It is typically used when a high-quality reference genome is available and represents the genome of interest accurately. The choice of strategy is determined by the specific research question and the available resources. *De novo* assembly enables the reconstruction of unique genomic sequences. Yet, it can be challenging and requires significant computational power. Mapping to a reference genome can reduce this computational cost and increase the accuracy of genome assembly. While mapping to a reference genome can be helpful, it may not adequately represent the genome of interest and may overlook novel genomic characteristics that don't appear in the reference genome (Lischer & Shimizu, 2017). For the strains isolated after 1940, *de novo* assembly was performed as the reference genome inaccurately represents the genome of interest, in contrast with the old strains.

The results of the *de novo* assembly show that for each of the three strains, the assemblies generated with SPAdes result in a larger genome length compared to the reference *P. aeruginosa* genomes. In addition, their GC content is lower than that of the references. The GC content is calculated by dividing it by the genome length. As the length differs between the assemblies and the references, their GC contents may differ as well. Another explanation could be that gaps and errors in the assemblies lower the GC content. Also, consider that small sequence differences between the assemblies and the references could affect the difference in GC content. Another difference is observed at the N50 level. Those of the assemblies are rather small compared to the references. According to OmicsBox (n.d.), higher N50 values indicate higher contiguity. This means that the assembly consists of fewer but larger sequences. This trend was to be expected. The references both consist of 1 scaffold (Table 4). This is confirmed by the fact that the N50 value of the scaffold is equal to the genome length. Since the assemblies consist of 13 and 15 scaffolds, they consist of smaller and more fragments and therefore have a lower N50 value than the references. The L50 statistic measures contiguity. The lower the L50 value, the fewer contigs or scaffolds are needed to cover 50% of the genome length. This means that there are fewer and longer sequences present. As expected,

the L50 values of references are smaller than those of assemblies because they both consist of only 1 scaffold (OmicsBox, n.d.).

Aminoglycoside resistance

The enzyme APH matches with all the strains, except for A22. This enzyme inhibits the mechanism of action of the antibiotic by transferring the γ-phosphoryl group of ATP to the 3'-hydroxyl end of the antibiotic (Kim *et al*., 2006). For the five strains, the E-value is 0. Because this parameter corresponds to the number of hits to be expected by chance when searching a library of this particular size, this enzyme is present in all new and two of the old strains. According to Pang *et al.* (2019), resistance through APH is part of bacterial intrinsic antibiotic resistance. This means that these conditions are thought to be innate. However, this is not uniform for all strains but is likely to be observed in most of them, which is confirmed by our own results. In fact, these enzymes are expected to be important for their own metabolism. In addition, it should be noted that genes encoding for aminoglycoside-modifying enzymes are known to be located on mobile elements, making them a subject for transposition (Poole, 2005).

For APH, no strong difference is seen between the old and new strains. For streptomycin 3'-adenylyltransferase, strain 1709-12, which is a new strain, shows 14 matches. This indicates that this strain may have multiple copies or variants of the gene encoding this enzyme. This enzyme transfers an AMP group of ATP to the 3' position of the added drug, making it inactive (Harwood & Smith, 1969). Because the enzyme is only noted in the new strain, indicating that this resistance is gained over time. This acquisition of antibiotic resistance is due to mechanisms such as mutation, gene transfer, and selection, and can occur over time in response to selection pressures, such as the use of antibiotics.

Fluoroquinolone resistance

Experiments performed by Nouri *et al*. (2016) showed that in *P. aeruginosa* the main mechanism of fluoroquinolones resistance is the Thr-83 to Ile substitution in gyrA. Other occurring amino substitutions are Asp-87 to Asn in GyrA and Ala-88 to Pro and Ser-87 to Leu in ParC.

For GyrA, SNPs are seen only in the new strains. In all three of the new strains, the amino acid substitution Thr-83 to Ile is detected, this is also the only substitution in strain Aa249. Strain 1709-12 and co380791 have 5 other substitutions. So for GyrA a clear difference can be seen between the old and new strains, only the new strains have mutations. So this indicates that the mutation was gained after the strains were in contact with the antibiotic.

Mutations in ParC are detected in both new and old strains, but not in all of them. The Ser-87 to Leu substitution is seen in new strains 1709-12 and co380791. In the new strain Aa249 and old strain A17 no SNPs were seen. For ParC the difference between the old and new strains is the amino acid substitution, the position of the SNPs are different.

Beta-lactam resistance

β-lactam is a commonly used antibiotic, for which resistance is rapidly rising. There are multiple resistance mechanisms against β-lactam. In *P. aeruginosa*, one of them is the overproduction of antibiotic inactivating enzymes (Pang *et al*., 2019). Juan *et al*. (2005) discovered that mutations, frameshifts and deletions, in *ampD* result in beta-lactam resistance. This resistance could be determined the same way the aminoglycoside and fluoroquinolones resistance was detected.

Conclusion

The purpose of the current study was to investigate the development of antibiotic resistance to aminoglycosides and fluoroquinolones in the bacterium *P. aeruginosa*. The second aim of this study was to explore the use of next-generation sequencing method Illumina HiSeq2500 and paired-end

sequencing data. Three strains isolated before, and three strains isolated after 1940, were compared to one another.

This study has found that the quality of raw sequencing data generated by Illumina HiSeq2500 is rather high and requires only little processing. The challenging nature of *de novo* assembly was revealed. While the *de novo* assemblies were of lower quality compared to the reference genomes, the obtained results are satisfactory.

One of the more significant findings to emerge from this study is that the aminoglycoside 3'-phosphotransferase, an AME protein, was present in two of the three old strains and remained unchanged in all the new strains. This, as well as the innate character of these enzymes (Pang et al., 2019), imply that antibiotic selection pressure is not the main cause for their presence. The second aminoglycoside-modifying enzyme is streptomycin 3’-adenylyltransferase. 14 matches were found in one of the new strains. This acquisition of resistance is evidence of selective pressure of antibiotic use.

For fluoroquinolone resistance, on the other hand, this study has found a large difference between new and old strains. Mutations in DNA gyrase subunit A, known to be present in antibiotic resistant strains, are only present in the new strains. Another important enzyme contributing to the fluoroquinolone resistance of *P. aeruginosa* is DNA topoisomerase IV (subunit A). Mutations in this enzyme are present in both old and new strains. Taken together, these findings suggest a significant role for selective pressure in promoting mutations in GyrA and streptomycin 3’-adenylyltransferase, but not in APH and ParC. Being limited to six *P. aeruginosa* strains, this study solely lays the groundwork for future research into antibiotic resistance in bacteria.

In conclusion, while resistance of *P. aeruginosa* to some antibiotics was already present in several strains before 1940, the findings reported here shed light on the increase of resistance development in *P. aeruginosa* over the last half-century. Antibiotic misuse and overuse represent selective forces for resistance development. These findings underscore the importance of responsible antibiotic usage and highlight the need for continued research to develop new strategies to combat antibiotic resistance.

References

Babraham Bioinformatics. n.d. Per Sequence GC Content. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html [retrieved: 19/03/2023]

Barnhart C, Campbel Rl, LaRosa LA, Marr AM, Morgan A, VanBerkom D. 2002. Mechanisms of Aminoglycoside Resistance. Retrieved from https://www.uphs.upenn.edu/bugdrug/antibiotic_manual/aminoglycosideresistance.htm [retrieved:

Batut, B., Doyle, M., Cormier, A., Bretaudeau, A., Leroi, L., Corre, E. & Robin, S. n.d. Quality Control. Retrieved from: https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#assess-quality-with-fastqc---short--long-reads [retrieved: 19/02/2023]

blast2go. n.d. FASTQ quality check. Retrieved from http://docs.blast2go.com/user-manual/tools-(pro-feature)/fastq-quality-check/#FASTQQualityCheck-SequenceLengthDistribution [retrieved: 17/03/2023]

CAB. n.d. QUAST manual. Retrieved from http://cab.spbu.ru/files/release2.2.0/quality.html

Chokshi A, Sifri Z, Cennimo D, Horng H. 2019. Global Contributors to Antibiotic Resistance. J Glob Infect Dis 11:36–42.

DHSS. n.d. What is Antibiotic Resistance. Retrieved from https://health.mo.gov/safety/antibioticresistance/generalinfo.php [retrieved: 15/03/2023]

DNA Technologies Core. n.d. PacBio Sequel II Library Prep & Sequencing. Retrieved from https://dnatech.genomecenter.ucdavis.edu/pacbio-library-prep-sequencing/

Egorov AM, Ulyashova MM, Rubtsova MY. 2018. Bacterial Enzymes and Antibiotic Resistance. Acta Naturae 10:33–48.
Harwood, J. H., & Smith, D. H. 1969. Resistance factor-mediated streptomycin resistance. Journal of Bacteriology, 97(3), 1262–1271. https://doi.org/10.1128/JB.97.3.1262-1271.1969.

Illumina. 2015. What Is De Novo Sequencing? Retrieved from https://emea.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing/de-novo-sequencing.html

Illumina. 2023. Performance specifications for the HiSeq 2500 System. Retrieved from https://www.illumina.com/systems/sequencing-platforms/hiseq-2500/specifications.html

Illumina. n.d. Advantages of paired-end and single-read sequencing. Retrieved from https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

Juan C, Maciá MD, Gutiérrez O, Vidal C, Pérez JL, Oliver O. 2005. Molecular Mechanisms of β-Lactam Resistance Mediated by AmpC Hyperproduction in *Pseudomonas aeruginosa* Clinical Strains. ASM Journals 49:4733-4738.

Kim, C., Joo, Y. C., Yan, H., Vakulenko, S. B., & Mobashery, S. 2006. Hydrolysis of ATP by aminoglycoside 3′ phosphotransferases: An unexpected cost to bacteria for harboring an antibiotic resistance enzyme. Journal of Biological Chemistry, *281*(11), 6964–6969.

Langendonk RF, Neill DR, Fothergill JL. 2021. The Building Blocks of Antimicrobial Resistance in *Pseudomonas aeruginosa*: Implications for Current Resistance-Breaking Therapies. Front. Cell. Infect. Microbiol 11.

Lischer HEL, Shimizu KK. 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18:474.

NCBI. n.d.a. Help for Assembly. Retrieved from https://www.ncbi.nlm.nih.gov/assembly/help/ [retrieved: 13/03/2023]

NCBI. n.d.b. Frequently Asked Questions: Q: What is the Expect (E) Value?. Retrieved from https://blast.ncbi.nlm.nih.gov/doc/blast-help/FAQ.html#faq [retrieved: 10/03/2023]

NCBI. n.d.c. ASM2549047v1. Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCA_025490475.1 [retrieved: 13/03/2023]
NCBI. n.d.d. ASM395710v1. Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCA_003957105.1#/qa [retrieved: 13/03/2023]
Nouri R, Rezaee MA, Hasani A, Aghazadeh M, Asgharzadeh M. 2016. The role of gyrA and parC mutations in fluoroquinolones-resistant Pseudomonas aeruginosa isolates from Iran. Brazilian Journal of Microbiology 47:925-930.

OmicsBox. n.d. Genome Assembly Quality Assessment. Retrieved from: https://manual.omicsbox.biobam.com/user-manual/omicsbox-modules/module-genome-analysis/genome-assembly-quality-assessment/.

Oxford Nanopore Technologies. n.d. MinION. Retrieved from https://nanoporetech.com/products/minion
PacBio. 2023. Genomes vs. genNNNes: the difference between contigs and scaffolds in genome assemblies. Retrieved from https://www.pacb.com/blog/genomes-vs-gennnnes-difference-contigs-scaffolds-genome-assemblies/

Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, Otto TD, Keane JA. 2016. Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data. Microb Genom 2.

Pang, Z., Raudonis, R., Glick, B. R., Lin, T. J., & Cheng, Z. 2019. Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. *Biotechnology Advances*, *37*(1), 177–192.

Poole, K. 2005. Aminoglycoside Resistance in Pseudomonas aeruginosa. Antimicrobial Agents and Chemotherapy, 49(2), 479. https://doi.org/10.1128/AAC.49.2.479-487.2005.

Skliros D. 2018. De Novo DNA Sequencing and the Special k-mer. Retrieved from https://bitesizebio.com/38250/de-novo-dna-sequencing-and-the-special-k-mer/

ThermoFisher. n.d. Ion GeneStudio S5 Systems. Retrieved from https://www.thermofisher.com/be/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing.html [retrieved: 28/02/2023]

Vicevall, E. 2017. What's N50? Retrieved from: https://www.molecularecologist.com/2017/03/29/whats-n50/ [retrieved: 19/02/2023]

Appendices

Appendix I: Overview of the strains analyzed in the current study

Table A.I.1: Data Old Strains (Origin, Sampling period, Antiobtoc resistance profile, Accession numbers of raw reads and assemblies)

OLD STRAINS						
Strain	ABR*	ID assembly	Source	Location		Sampling period
A15	1	GCF_003837025.1	Wound	Paris	France	1882-1918
A17	1	GCF_003836875.1	Leg ulcer	Paris	France	1882-1918
A22	1	GCF_003836865.1	Wound	Paris	France	1882-1918

Table A.I.2: Data New Strains (Origin, Sampling period, Antiobtoc resistance profile, Accession numbers of raw reads and assemblies)

NEW STRAINS						
Strain	ABR*	ID raw data	Source	Location		Sampling period
Aa249	5	SRR2939461	Blood	Aachen	Germany	1997
co380791	4	SRR2939512	Blood	Cali	Columbia	2003
1709-12	5	SRR2939511	clinical non CF	Leuven	Belgium	2004

Appendix II: Quality assessment in Galaxy (Batut *et al.,* 2022)

1. Per base sequence quality

This data is represented in a plot with the x-axis showing the position in the read and the y-axis the quality score (Phred score). For each position in the read, a boxplot is drawn. The 25%-75% range is represented by a filled yellow box. The upper and lower whiskers represent the 10% and 90% values respectively. The red line is the median value. The blue line is the mean quality.

The background of the plot exists of three colors dividing the plot into a zone with ‘very good quality scores’ (green), ‘reasonable quality scores’ (orange), and ‘poor quality scores’(red). High-quality data is in the green zone. Additionally, the quality at the end of the read should stay high.

2. Per sequence quality scores

This data is represented in a plot with the number of reads on the y-axis and the mean sequence quality score (Phred score) on the x-axis. In high-quality data, there should be a tight peak at high mean sequence quality scores. A peak at smaller values indicates some sequences were sequenced with lower quality. These peaks are undesirable, and if present should represent only a small percentage of the total data.

3. Per base sequence content

This data is represented in a plot with the percentage value of the base (T, C, A, G, each with a distinct color) on the y-axis and the position in the read on the x-axis. In high-quality data, there should be little or no difference between the occurrence of all four bases, as well as an approximatively constant percentage of each base over the entire read. Additionally, %A should be equal to %T, and %C equal to %G.

4. Per sequence GC content

This data is represented in a plot with the number of reads on the y-axis and the mean GC content on the x-axis. A theoretical reference distribution is calculated from the observed data and shown in blue. The central peak in this uniform distribution is equal to the overall GC content of the underlying genome. The actual GC content per read is shown in red.

In bad Illumina data, undesirable unusually-shaped distributions are observed. This is due to contamination of the library or some kind of bias in the data.

5. Per base N content

This data represents the proportion of positions in the read for which no base was called. An N indicates that the sequencer was not able to call a base with sufficient confidence at that specific position in the read. For high-quality data, this value should be very low or equal to zero. High proportions of N imply a loss of sequencing quality.

6. Sequence Length Distribution

This plot shows the distribution of sequence lengths over all sequences in the analysis. Often the plot will have a single peak at a specific sequence length. If the file contains fragments of variable length, different peaks may appear with the height of the peak representing the relative amount of sequences of that length.

7. Sequence Duplication Levels

This data is represented in a plot with the percentage of sequences on the y-axis and the sequence duplication level on the x-axis. Duplication can occur during PCR, but can also be due to an actual biological overrepresentation of a sequence. A biased enrichment in PCR of specific sequences will lead to misrepresentation of the proportion of those sequences in the file.

8. Overrepresented sequences

In libraries with diverse sequences, overrepresentation is not expected. Thus, the overrepresentation of a specific sequence could indicate contamination. It could also indicate actual biological overrepresentation, implying the importance of that sequence.

9. Adapter Content

In this plot, the percentage of different adapters in the reads is analyzed for all the positions in the read. The occurrence of the adapters is displayed in a cumulative manner. Once an adapter sequence is observed, it is counted as present through to the end of the read. Therefore the percentage of that adapter increases with the read length.

High-quality data should not have any adapter present. For sequences shorter than the read length, there will be read-through activity at the 3’ end of the sequence, resulting in the presence of the adapter sequence in the data.

Appendix III: Parameters used in Trimmomatic

Perform initial ILLUMINACLIP step?	yes
Select standard adapter sequences or provide custom?	standard
Adapter sequences to use	TruSeq3 (paired-ended, for MiSeq and HiSeq)
Maximum mismatch count which will still allow a full match to be performed	2
How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment	30
How accurate the match between any adapter etc. sequence must be against a read	10
Minimum length of adapter that needs to be detected (PE specific/palindrome mode)	8
Always keep both reads (PE specific/palindrome mode)?	True
Select Trimmomatic operation to perform	HEADCROP
Number of bases to remove from the start of the read	5
Output trimlog file?	False
Output trimmomatic log messages?	False
Job Resource Parameters	no

Figure A.III.1: Parameters used in Trimmomatic

Appendix IV: Visual Representation of the N50 and L50 parameters

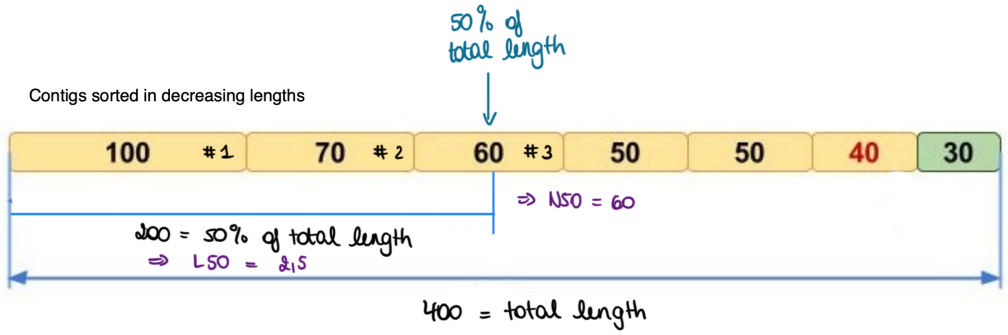

Figure A.IV.1: Visual representation of the N50 and L50 values. Figure adapted from The Molecular Ecologist (2018).

Appendix V: Quast

Figure A.V.1: Heatmap used in Quast to compare the quality of the contig and scaffold outputs.

Statistics without reference	Galaxy5-__ob__SPAdes_on_data_...	Galaxy6-__ob__SPAdes_on_data_...
# contigs	150	112
# contigs (>= 0 bp)	250	212
# contigs (>= 1000 bp)	141	104
Largest contig	404 353	674 638
Total length	7 080 024	7 081 048
Total length (>= 0 bp)	7 105 674	7 106 698
Total length (>= 1000 bp)	7 073 418	7 075 046
N50	115 197	158 361
N90	30 017	49 479
auN	144 745	215 581
L50	19	13
L90	64	44
GC (%)	65.85	65.85
Mismatches		
# N's per 100 kbp	0	15.53
# N's	0	1100

Figure A.V.2: Quality assessment of the genome assembly of Aa249. Left: Quast on the contigs generated in SPAdes. Right: Quast on the scaffolds generated in SPAdes.

Statistics without reference	Galaxy8-__ob__SPAdes_on_data_...	Galaxy9-__ob__SPAdes_on_data_...
# contigs	182	140
# contigs (>= 0 bp)	523	480
# contigs (>= 1000 bp)	153	112
Largest contig	373 509	662 617
Total length	7 228 486	7 229 660
Total length (>= 0 bp)	7 291 166	7 292 418
Total length (>= 1000 bp)	7 207 486	7 209 197
N50	104 008	177 679
N90	26 938	41 612
auN	128 580	220 317
L50	21	13
L90	72	43
GC (%)	65.66	65.66
Mismatches		
# N's per 100 kbp	0	18.26
# N's	0	1320

Figure A.V.3: Quality assessment of the genome assembly of Co380791. Left: Quast on the contigs generated in SPAdes. Right: Quast on the scaffolds generated in SPAdes.

Statistics without reference	Galaxy9-__ob__SPAdes_on_data_...	Galaxy10-__ob__SPAdes_on_data...
# contigs	181	154
# contigs (>= 0 bp)	448	419
# contigs (>= 1000 bp)	153	128
Largest contig	311 029	399 153
Total length	7 050 650	7 051 764
Total length (>= 0 bp)	7 092 190	7 093 144
Total length (>= 1000 bp)	7 029 749	7 032 389
N50	118 419	154 612
N90	27 947	36 373
auN	118 967	171 241
L50	22	15
L90	70	51
GC (%)	65.83	65.83
Mismatches		
# N's per 100 kbp	0	15.6
# N's	0	1100

Figure A.V.4: Quality assessment of the genome assembly of 1709-12. Left: Quast on the contigs generated in SPAdes. Right: Quast on the scaffolds generated in SPAdes.

Appendix VI: Protein Blast Search and Multiple Sequence Alignment

Table A.VI.1: Blastp of the AME file against the generated protein database. Columns 1: Query Seq-id, 2: Subject Seq-id, 3: Percentage of identical matches, 4: Alignment length, 5: Number of mismatches, 6: E-value. 7: Protein.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
ag:CAA26199	1709-12_05732	87.109	256	33	6.89e-166	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:AAB59083	1709-12_05732	86.719	256	34	6.52e-166	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:CAA48308	1709-12_05732	100.000	259	0	0.0	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:AAC14728	1709-12_05732	98.069	259	5	0.0	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:AAK69614	1709-12_05732	93.822	259	16	3.11e-180	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:CAJ13568	1709-12_05732	94.981	259	13	0.0	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:AAV49760	1709-12_05732	94.208	259	15	2.72e-180	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:AFQ94781	1709-12_05732	83.398	259	43	1.19e-158	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:ABD58917	1709-12_05732	92.188	256	20	1.23e-174	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:ACK43806	1709-12_05732	94.981	259	13	0.0	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:AAN87151	1709-12_05732	88.803	259	29	4.74e-171	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:CAK12750	1709-12_05732	88.803	259	29	4.95e-172	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:CAH10847	1709-12_05732	89.575	259	27	9.58e-173	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:CAM98048	1709-12_05732	87.645	259	32	1.63e-167	1709-12_05732 Streptomycin 3"-adenylyltransferase
ag:CAA62365	1709-12_06126	98.507	268	4	0.0	1709-12_06126 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	co380791_04948	98.507	268	4	0.0	co380791_04948 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	Aa249_02951	98.507	268	4	0.0	Aa249_02951 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	A17_05014	98.507	268	4	0.0	A17_05014 Aminoglycoside 3'-phosphotransferase
ag:CAA62365	A15_01328	97.388	268	7	0.0	A15_01328 Aminoglycoside 3'-phosphotransferase

Table A.VI.2: Blastp of ParC against the generated protein database. Columns 1: Query Seq-id, 2: Subject Seq-id, 3: Percentage of identical matches, 4: Alignment length, 5: Number of mismatches, 6: E-value. 7: Protein.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
NP_253651.1	Aa249_00028	100.000	754	0	0.0	Aa249_00028 DNA topoisomerase 4 subunit A
NP_253651.1	A17_05712	100.000	754	0	0.0	A17_05712 DNA topoisomerase 4 subunit A
NP_253651.1	A15_02282	99.867	754	1	0.0	A15_02282 DNA topoisomerase 4 subunit A
NP_253651.1	A22_03214	99.867	754	1	0.0	A22_03214 DNA topoisomerase 4 subunit A
NP_253651.1	1709-12_05146	99.867	754	1	0.0	1709-12_05146 DNA topoisomerase 4 subunit A
NP_253651.1	co380791_00027	99.867	754	1	0.0	co380791_00027 DNA topoisomerase 4 subunit A

Table A.VI.3: Blastp of GyrA against the generated protein database. Columns 1: Query Seq-id, 2: Subject Seq-id, 3: Percentage of identical matches, 4: Alignment length, 5: Number of mismatches, 6: E-value. 7: Protein.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
ref|NP_251858.1|:1-923	A22_04912	100.000	923	0	0.0	A22_04912 DNA gyrase subunit A
ref|NP_251858.1|:1-923	A17_00177	99.784	925	0	0.0	A17_00177 DNA gyrase subunit A
ref|NP_251858.1|:1-923	Aa249_06310	99.676	925	1	0.0	Aa249_06310 DNA gyrase subunit A
ref|NP_251858.1|:1-923	A15_01936	99.783	923	0	0.0	A15_01936 DNA gyrase subunit A
ref|NP_251858.1|:1-923	1709-12_04977	99.133	923	6	0.0	1709-12_04977 DNA gyrase subunit A
ref|NP_251858.1|:1-923	co380791_06469	99.133	923	6	0.0	co380791_06469 DNA gyrase subunit A

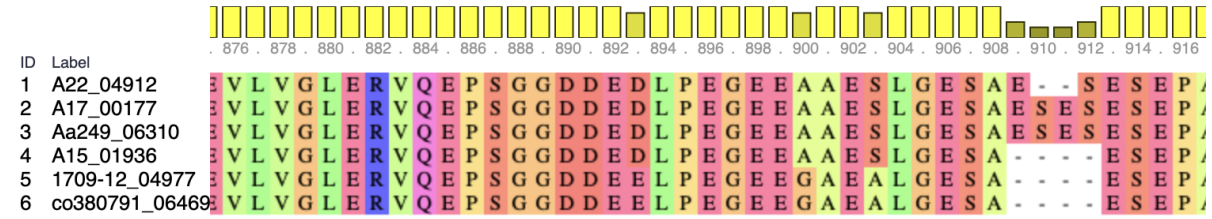

Figure A.VI.1: Multiple sequencing alignment of the GyrA protein sequences obtained in the blast search. The yellow bars on the top indicate the sequence conservation.

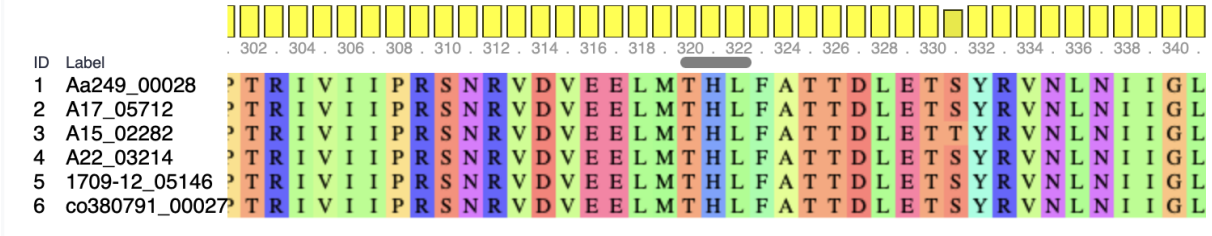

Figure A.VI.2: Multiple sequencing alignment of the ParC protein sequences obtained in the blast search. The yellow bars on the top indicate the sequence conservation.