VERSION 14 June, 2016

# GENOMIC RESISTANCE GENE DETECTION, INTEGRON DIVERSITY MAPPING AND RESISTOME-WIDE ASSOCIATION ANALYSIS HELP DEFINE ANTIBIOTIC RESISTANCE IN *PSEUDOMONAS AERUGINOSA*

**Magali Jaillard[1], Alex van Belkum[1*], Kyle Cady[2], David Creely[1], Dee Shortridge[1], Bernadette Blanc[1], Magda Barbu[2], W. Michael Dunne Jr[1], Gilles Zambardi[1], Mark Enright[3], Nathalie Mugnier[1], Pierre Mahé[1], Christophe Le Priol[1], Stéphane Schicklin[1], Ghislaine Guigon[1] and Jean Baptiste Veyrieras[1]**

[1]bioMerieux SA, 376, Chemin de l'Orme, 69280 Marcy l'Etoile FRANCE
[2]Synthetic Genomics, Inc., 11149 North Torrey Pines Road, La Jolla, CA 92037, USA
[3]Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK.

[*]**Communicating author:** Alex van Belkum

**Address for correspondence:** bioMérieux, Microbiology R&D,
3 Route de Port Michaud,
38390 La Balme Les Grottes, France
e-mail alex.vanbelkum@biomerieux.com
phone 0033609487905.

**Conflicts of interest:** All authors except ME are (part-time) employees of bioMérieux or Synthetic Genomics Inc and hence have a business implication in all work presented here. However, the study was designed and executed in an open manner and all data generated have been deposited in the public domain, also resulting in the current publication.

## Abstract

**Objectives:** Genetic determinants of antibiotic resistance have been extensively investigated in the past using relatively small panels of clinical microbial isolates. Today, next generation sequencing allows a broader assessment of the links between genotype and phenotypic antibiotic susceptibility. Such large-scale association studies are expected to refine the present understanding of the genetics of antibiotic resistance.

**Methods:** We analyzed 672 strains of *Pseudomonas aeruginosa,* including representatives of the major multi (MDR) and extensively (XDR) drug resistant clones, for which genome sequences and multiple parameter antibiograms were available. We correlated variability in antibiograms and minimal inhibitory concentrations (MICs) for those antibiotics with the occurrence of resistance genes, performed a search for and characterization of integrons and developed a resistome wide association study (RWAS) for the nine most relevant and clinically useful antibiotics. We distinguished intrinsic and extrinsic resistance genes and single nucleotide polymorphisms and insertions-deletions (indels) in such genes while controlling for potential confounding effects due to population structure

**Results:** The complex resistome of *P. aeruginosa* was visualized at the population level and gene prevalence could be defined at the species level. The previously assessed intrinsic genes were identified and a large number of more mobile and less frequent genes were mapped. A visual representation of combined genotypic and phenotypic data was developed.

Using a bioinformatics approach, integron content and variation was established showing 28 non-unique integron types carrying a variety of antibiotic resistance genes. Some of the integrons closely associated with known clonal complexes whereas others were more randomly distributed across the phylogenetic tree..

The RWAS led to the correct identification of the main known resistance mechanisms for meropenem, amikacin, levofloxacin and cefepime, but also to the detection of 46 new mutations associated with an increase or decrease of MIC for meropenem, amikacin, cefepime and colistin antibiotics. Among these new candidates, 29 were variants of the *opr*D gene and these associated with variation in the meropenem MIC.

**Conclusions:** The results shown here for *P. aeruginosa* validated different statistical association frameworks based on which MIC and resistome-wide data could be correlated. We developed a system for integron mapping and characterization using genome sequences. The same approaches can be applied to other microbial species ultimately facilitating the prediction of an antibiogram on the basis of genomic resistance gene content.

**Key words:** Antibiotic resistance – *Pseudomonas aeruginosa* – RWAS – integrons – resistance gene mapping

## Introduction

*Pseudomonas aeruginosa* strains have large genomes with a degree of plasticity allowing adaptation to a diversity of pathogenic and natural environments. The species is genetically variable due to efficient recombination and acquisition of mobile genetic elements (MGEs) [42]. *P. aeruginosa* absorbs exogenous DNA and under selective antibiotic pressure it efficiently gathers resistance traits [33, 43]. Accumulation of resistance traits is more likely to occur in isolates with enhanced virulence [3, 34], during infection and treatment, and hence *P. aeruginosa* selectively evolves resistance to antibiotic-mediated killing [10]. In spite of the significant genetic variability within the species, resistant clones that spread globally have been identified [103, 104, 105]. Among them, multi- and extensive- drug resistant (MDR and XDR) clones are common, often related to the expansion of the content of integrons [44]. These pandemic clones include the highly prevalent sequence types ST235, ST111, ST348 and ST175 [13, 14, 17, 35].

In the field of antibiotic resistance (AR) and susceptibility testing (AST) [38], progress needs to be made for better delineation of existing and novel AR mechanisms and resistance genes [106]. Genome sequencing of *P. aeruginosa* libraries identified molecular markers for resistance to amikacin, meropenem and levofloxacin [30]. Additionally, the international *P. aeruginosa* consortium published a summary of the species' resistome, recognizing 73 known AR genes in 389 isolates and highlighting the importance of the accessory genome in overall antibiotic resistance [41]. Our recent studies revealed that there may be differences in the resistome of strains equipped with different CRISPR-Cas bacterial immune systems [99]. However, CRISP-Cas immunity does not seem to be directly blocking the acquisition of resistance elements, although the CRISPR spacers do possess a small number of target sequences against known plasmids that may carry resistance elements themselves [99]. Hence, additional studies on resistance gene content of the *P. aeruginosa* pan-genome are still much needed.

Studies for other pathogens have suggested that genomic antibiograms can perform as well as phenotype-based techniques. The first large study in this field documented 99.7% concordance between geno- and phenotypic antibiograms [27]. Subsequent work on *Staphylococcus aureus*, *Escherichia coli*, *Klebsiella pneumoniae* and *Campylobacter* spp confirmed and extended these findings [28, 29, 108], whereas for

a strictly clonal bacterial species such as *Mycobacterium tuberculosis* very impressive results were obtained [107]. A recent resistome analysis for *P. aeruginosa* correlated meropenem resistance with outer membrane protein OprD polymorphism but other factors including metallo-beta-lactamases and *Klebsiella pneumonia* carbapenemases played important roles as well [30]. Levofloxacin resistance was found to primarily depend on QRDR variation in DNA gyrase and topoisomerase mutations. Amikacin resistance was mostly associated with efflux systems and the presence of aminoglycoside modifying enzymes. Clearly, genomic approaches hold promise for the development of future AST systems for routine use in clinical microbiology laboratories, although knowledge gaps still need to be filled. In order to precisely predict resistance or susceptibility from a genome minimum standards for genome coverage, assembly and annotation need to be agreed upon prior to its widespread use in routine diagnostics [109].

Here we present a genome-based resistome characterization, a descriptive study of integron dynamics and a resistome-wide association study for 672 *P. aeruginosa* strains for which the complete genome sequence was known and for which MIC values for the 9 most often used anti-*P. aeruginosa* antibiotics were available.

## Materials and methods

**Description of the strain panel:** Genomic information for 672 *P. aeruginosa* strains was analyzed. Strains were obtained from 3 diverse collections: the bioMérieux collection (n=219) [99], the Kos collection (n=390) (30), and the Pirnay collection (n=63) [110]. Strains were identified to the species level based on the use of VITEK2 ID 32 A identification cards and API analyses and, if needed, specific sequencing of 16S rRNA and EF-Tu genes (results not shown). Genome sequences were obtained as previously reported [30, 99, 110] and assembled and annotated for the identification of core and variable genes. Genomic MLST and phylogenetic comparisons were performed. For the strains in the Pirnay and bioMérieux collections, reference antibiotic resistance data based upon broth dilution assays complemented with VITEK2 data were available for more than 20 antibiotics for most strains. In addition, a public database including resistance testing and genome sequences was retrieved for 388 *P. aeruginosa* isolates of the Kos collection since

two isolates were not tested for antibiotic resistance [30]. Resistance data included results for meropenem, amikacin and levofloxacin. More detailed information on all of the strains and their analyses can be obtained from recently published work [99]. We have now developed a database including information on the nine antibiotics included in this study (Supplementary Table XX??).

**Resistome:** The resistome of the 672 *P. aeruginosa* isolates was obtained by annotating each genome assembly versus an in-house pan-bacterial antibiotic resistance sequences database. This database contains 2545 non-redundant reference sequences categorized into 569 antibiotic resistance (AR) genes. More specifically, among these reference sequences, the database contains at least one sequence for each previously reported AR gene including genes directly involved in susceptibility and/or resistance *per se* in *P. aeruginosa* [33].

Each genome was then searched via BLASTn (Version 2.2.28+) for resistance genes with at least 80% identity and at least 80% coverage relative to at least one gene within our 569 gene library. Overlapping alignments belonging to the same AR gene were clustered into a single hit if the overlap covered more than 10% of the aligned reference sequences. For each cluster, we only kept the best hit defined as the alignment with the highest percentage identity times the reference sequence coverage. This best hit was retained to infer the haplotype state of the gene (e.g. if the best hit was obtained with the reference sequence of OXA-31, we annotated the gene as an OXA-31 variant). Note that for a given isolate and for a given AR gene several distinct annotations (i.e. distinct copies) can be reported when hits appeared at distinct genomic locations. Supplementary Table 5 reports the genotypes (copy number of genes and allelic doses of mutations) of each of the 147 determinants that were found at least once among the 672 genomes.

**Integron analyses:** We detected integrase genes by aligning (tblastn) the assembled genome sequences (in all reading frames) with the protein sequences of the three known integrases: intI1, intI2 and intI3 as derived from a targeted UniProt search. The integrase sequences used to annotate the strains are provided in Supplementary File 1. Contrary to intI1 for which we were able to retrieve 36 reference sequences, we found only 3 reference sequences for intI3 and only 1 reference sequence for intI2. We used the same strategy that was used for the antibiotic resistance gene annotation with a 70% identity and 70% protein coverage cutoff. We then searched for recognizable and annotated AR genes upstream of the

integrase start codon and reported the physical distance between the integrase start codon and the 5' end of the resistance element. We included all resistance elements within a 10 kb window thereby assuming that all resistance elements being part of the integron would thus be found.

**Resistome genotyping:** Based on our resistome annotations, we genotyped the 672 isolates for allele counts at both the "locus-level" and the "variant-level". By allele counts at the locus level we mean the distinct number of copies of a given AR gene (e.g. TEM, mexX, etc) irrespective of the alleles of the gene. Regarding the allele counts at the variant-level, we aligned all the detected sequences of a given AR gene using mafft (version 6.861 using high-precision mode see reference [100]) in order to identify both SNPs including tri- and quadri-allelic sites and indels. Finally, for a given isolate, we counted the number of occurrences for each SNP and indel allele and used this as the corresponding allele count at the variant-level.

**Resistome-wide association study (RWAS):** Our objective here was two-fold: we first assessed the strength of association between causal genotypes known to increase (or decrease) the MIC value (generally beyond the resistance breakpoint) and, second, we looked for novel candidate genotypes that may be associated with an additional variability of MIC values taking the effect of known genotypes into account. Figure 1 provides a literature-derived overview of the causal genes or variants thereof known to increase (or decrease) the MIC value for each antibiotic in our association study.

**Population structure:** Association results may be inflated by the presence of cryptic correlations between population structure and the MIC status of the strains. This could be due to either population-wide linkage disequilibrium (LD) between the causal mutations and the genetic structure, or to a sampling bias leading to over-representation of related strains. We prevented confounding effects due to population structure by computing the principal components (PC) from the isolate covariance matrix derived from the non-resistome related core-SNP genotypes [53]. By "non-resistome related" we mean that we removed SNPs from core-genes that are also part of our resistome database (gyrA, gyrB, parC, parE, folP, OXA-50, ampC, etc) and that were subsequently evaluated for association with the MIC values. The PC scores of the isolates have been used as covariates in the statistical association framework.

**Associative modeling:** For each antibiotic we used ordinal regression [45] to develop the core statistical association framework. This provides the benefit of adequately accounting for the ordinal nature of the MIC values and it facilitates inclusion of covariates into the association model, in order to both control for population structure, and to test for the additional predictive value of genotypes with respect to known causal genotypes.

For each antibiotic we first forwardly included any principal components (PC) with a p-value of 0.05 or less as covariates into the ordinal regression model. We thus built Z, the n x k matrix of PC where n is the number of isolates with a non-missing MIC value and k the number of retained PC after the forward inclusion procedure. After the inclusion of the PC as covariates, we added all known genotypes to the retained model. Considering only known genotypes as eligible, we then performed a backward elimination procedure based on the Akaike Information criterion (AIC) and stopped when any removal of a known genotype increased the AIC with respect to the previous model [101]. By doing so, we intended to select the optimal subset of known causal genotypes associated with MIC variability while preserving statistical power for detecting new associations. This step led to a matrix U of size n x p, of genuine causal genotype counts where p is the number of genotypes retained after the backward elimination procedure. Finally, we evaluated novel candidate genotypes for locus- and variant-allelic counts. These tests were performed independently for each candidate genotype x. If we denote Yi the MIC values for isolate i=1..n assuming that Yi can fall into j=1..J ordered categories (from the lowest tested antibiotic concentration to the higher one), we then considered the following cumulative link model for each candidate $P(Yi \leq j) = \pi i1 + \ldots + \pi ij$, where πij for j=1..J denotes the probability that the ith observation falls in the jth MIC category. Our cumulative logit regression model is then:

$$\begin{aligned} \text{logit}[\,P(\boldsymbol{Y_i} \leq \boldsymbol{j})\,] &= \log\left[\frac{P(\boldsymbol{Y_i} \leq \boldsymbol{j})}{\mathbf{1} - P(\boldsymbol{Y_i} \leq \boldsymbol{j})}\right] \\ &= \theta_j - (Z\alpha + U\beta + x\gamma) \end{aligned}$$

where : { Θj } for j=1..J stands for the MIC category intercepts, α is the vector of k regression parameters for the PC scores capturing the genetic structure, β is the vector of p regression parameters for the known causal genotypes, and γ is the regression parameter for the candidate genotype x.

To compute the p-value of the association between the candidate genotypes and the MIC variability, we used the standard likelihood-ratio test (LRT) [102]. This test compares the log-likelihood of the ordinal regression model under the null hypothesis (H0: the genotype has no effect on MIC variability; γ = 0) versus the alternative hypothesis (H1: the genotype has an effect on MIC variability; γ ≠0). If we denote L0(Θ, α, β; Y) and L1(Θ, α, β, γ; Y)) the log-likelihoods under each hypothesis, we can derive the likelihood-ratio test formula: -2 [ L1 – L0 ] ~ $|^2$ and compute the association p-value of each tested genotype from this formula.

Finally, we reported significant associations by controlling the false-discovery rate for each antibiotic at 5% using the Benjamini and Hochberg procedure [46]. We used the R package ordinal (R Development Core Team, 2011) to implement our cumulative logit regression framework.

## Results

**Global resistome and association analysis for known resistance markers:** We assembled antibiogram data for 672 isolates, however, our analysis focused only on the 9 most clinically relevant drugs belonging to 6 main families: aminoglycosides (amikacin), beta-lactams (meropenem, cefepime and piperacillin), quinolones (levofloxacin), polypeptides (colistin and polymyxin B), fosfomycins (fosfomycin) and phenicols (chloramphenicol) (Supplemental Table 4). Resistome analysis identified 147 loci previously suspected or confirmed to provide resistance to these major antibiotic families (see Figure 1, left column) . While most resistance elements were found to be extra-chromosomal, 45% of AR elements were chromosomally encoded. Note that we did not specifically address locations on MGEs or plasmids. The 147 AR elements cover 5 mechanisms to neutralize antibiotic action, involving both intrinsic and acquired resistance traits. Figure 1 also shows the population structure, phenotypic resistance markers and genome size variability and confirms that

resistance related to drug efflux and the activity of enzymatic pumps is mainly intrinsic. Traits dealing with antibiotic inactivation are mainly acquired.

**Integron analyses:** Integrase genes were found in one third of isolates (n=229), and an AR gene was identified upstream of two thirds of these integrase motifs (n=163). Interestingly, only class 1 integrons were detected in the current study. Thirty-seven different resistance genes were found to be co-localized with the integrase, and were therefore identified as being integron associated. OXA and *sul*1 were found to be the most frequent AR genes: among the 163 integrons carrying AR genes, only 17 did not carry OXA nor *sul*1 genes. Class I integrons are known to carry a large variety of resistance genes, particularly those targeting phenicol, aminoglycoside or beta-lactam antibiotics [42, 43]. In particular, there are known to encode aminoglycoside modifying enzymes, adenylytransferases and acetyltransferases [39, 40, 42, 43]. We found these gene families to be enriched in the *P. aeruginosa* integrons, as shown by the Poisson distribution test result in Figure 2A: enrichment is defined as a higher proportion of gene found in integron compared to other genome location, and is shown by a star. Over these 14 genes which are significantly more frequent within the 10kb upstream region of int1, 11 belong either to the adenylytransferase (*aad*) or acetyltransferase (AAC) gene families

Seventy-two different AR gene combinations were identified in the integrons of which the most prevalent one occurred 27 times and only contained the *sul*1 and *aad*A6 AR genes. This cassette was identified as the In51 integron [39], containing, ordered from the integrase *Intl*1: *aad*A6, *orf*D, *qac*EΔ1 and *sul*1 genes. Three isolates carried a truncated form of In51, with either a deletion within the antiseptic resistance gene *qac*EΔ1 or complete deletion of the *orf*D gene, as shown in Figure 2B.

Figures 2C and 2D highlight the great diversity of integron cassettes, by showing the AR gene combinations found more than 3 times (Figure 2C) and their variable distribution along the phylogenetic tree (Figure 2D). Thus we can observe integrons which are specific to small clades, such as the one carrying AR cassettes #2, #3, #14 or #15 and integrons that spread among all phylogroups, such as the one carrying AR cassettes #1 (In51), #4 or #7. Obviously, such analyses identify physically-associated resistance traits and generate evidence of horizontal transfer between distantly related strains.

**Correlating geno- and phenotypes through RWAS:** We carried out a detailed association study for the 9 major clinically used antibiotics and the combined

phenotypes (MIC values) and genotypes (copy number of genes (Figure 3A) and allelic doses of mutations - both SNPs & INDELs - within each gene (Figure 3B)) for all isolates using ordinal regression. Results are summarized in Figure 3 where both known and new associations are reported. Around 15% of the known determinants were not found to be associated because they were only present once or twice among all isolate genomes. Our test had not enough statistical power to detect these variants with extremely low allele frequencies. This concern 17 genes, shown on the left side of Figure 3A (the occurrence number within the 672 genomes is given below each gene column) and some mutations in *gyrA* and *gyrB* genes, e.g. *gyrB*:467/468 is present only once in the panel.

Very low prevalence of a particular resistance status also prevents from detecting significant associations. In Figure 3, antibiotic rows are ordered by decreasing prevalence; polymyxin B, colistin and chloramphenicol all present prevalence of non-sensitive isolates below 3%: 3 non-sensitive among 117 isolates for polymyxin B, 15 non-sensitives among 524 isolates for colistin and 3 sensitives among 103 isolates for chloramphenicol (see Supplementary Table 1). This may explain the absence of significant associations between known genotypes and MIC values for these 3 antibiotics. However, thanks to a large sample size available for colistin (n=524), conversely to the low sample size for polymyxin B and chloramphenicol, we found two new associations for this drug. First, the absence of the *mexS* gene (efflux pump) in 24 isolates yields an increased susceptibility: 50% of the isolates without mexS exhibit a MIC value less than 0.5ug/ml. This is in contrast with recent analyses where no such correlation was found [111]. Second, a deletion of the 5'end of the *ampD* gene (which varies in size in our panel up to 70 bp from the 5'end, although to avoid redundancy we reported it as a single indel in Figure 3) observed in 12 isolates seems to be related to an increase of the MIC values of these isolates (60% of them have a MIC value greater than or equal to 2 g/ml, with respect to 3% for the 512 isolates with the non-deleted version of *ampD*, see Supplementary Figure 4).

The lack of statistical power due to a limited number of isolates with MIC values may explain the negative results for fosfomycin (n=113 and 36 resistant isolates). More surprisingly, no significant association was observed for piperacillin (n=280 and 114 resistant isolates). There is no strong population structure that could explain these poor results (see Supplementary figure 11).

For cefepime, we confirmed the association with the presence of OXA but due to the low frequency of the gene AER (2 occurrences in our panel), we were unable to further test the association between the latter gene and cefepime susceptibility. Interestingly, we found cefepime resistance associations with OXA mutations which suggest that specific alleles of OXA correlate with higher MIC values (in particular mutations related to haplotypes OXA-31 and OXA-224, see also Supplementary Figure 5) [112].

Most of the known resistance genotypes for levofloxacin were found to be significantly associated with the MIC values (except the presence of AAC(6')-Ib3, and some mutations in *gyrA* and *gyrB*, for which the frequencies in our panel were too low). The two common non-synonymous mutations *gyrA*:T83I (38% of the isolates) and *parC*:S87L (26% of the isolates) show the highest association scores (p-value = $1.29x10^{-61}$ and $3.52x10^{-40}$, respectively). No additional association was detected.

Regarding amikacin, 9 out of the 26 genes (mostly the AAC(6') family) known to confer resistance were found to be statistically significantly associated with elevated MICs. Among the 17 undetected genes, 7 show extremely low allele frequencies. Regarding new associations, isolates lacking the efflux pump genes, mexX (10 isolates) and mexY (14 isolates), tend to exhibit lower MIC values: in both cases around 50% of the *mexX*/*mexY* deleted isolates have MIC values lower or equal to 0.5ug/ml (see Supplementary Figure 2A). Surprisingly, *gyrA*:T83I absence or presence is strongly associated with the amikacin MIC values, suggesting either a joint selective pressure for resistance to levofloxacin and amikacin or a sampling bias in our panel where closely-related MDR isolates could be over-represented. The association with the presence of OXA gene is likely also to be an artifact since OXA genes are a proxy of resistance cassettes linked to class I integrons (as discussed previously). This may suggest the presence of additional resistance genes (e.g. rare variants of AAC(6') genes) within the integron cassette not adequately detected by our current annotation pipeline.

Six out of the 23 known genotypes for meropenem resistance showed a statistically significant association, in particular for the presence of beta-lactamases VIM, VEB, OXA, KPC and IMP (whatever the allelic state of these genes). Here again, 7 out of the 17 undetected genes were not detectable due to extremely low allele frequencies. Despite high allele frequencies we did not document significant associations for CARB, GES and PER. The absence of the porin gene *oprD* (about

4% of the tested isolates) was also confirmed as significantly associated with increase in MIC values: 75% of the isolates lacking *oprD* show MIC values ≥ 8 when only 25% of isolates with *oprD* reach such value (see Supplementary Figure 8a). Regarding new associations for meropenem, *gyrA* and *parC* non-synonymous mutations related to levofloxacin resistance were also identified. The same assumptions as for amikacin resistance associated to *gyrA* mutation may apply here. Interestingly, several mutations within the porin gene *oprD* showed a significant association signal with meropenem MIC variability. Most of them are in strong linkage-disequilibrium (see Figure 3): by clustering mutations with a LD of at least 0.8 (as measured using *r*-squared), we can distinguish 6 groups among which 3 singletons and 1 group with only two linked mutations. The first singleton is the mutation leading to an early-stop codon at the beginning of the gene (*oprD*:Q18* in Figure 3) carried by 38 isolates in our panel for which the minimal MIC value was 0.25ug/ml and 60% of the 38 isolates have a MIC >8ug/ml.

Another early-stop codon, *oprD*:K350*, a singleton slightly linked to *oprD*:Q18*, was detected at the end of the gene and observed in 28 isolates, leading to a similar increase of the MIC values as for *oprD*:Q18*.

Conversely, the two linked synonymous mutations (*oprD*_180_c and *oprD*_758_a in Figure 3) are associated with more susceptible isolates (60% of them have a MIC value ≤ 0.25 while this proportion falls to 25% for alternative genotypes). The fact that these mutations do not result in changes to the protein sequence may suggest that they are in LD with a non-identified causal mutation elsewhere.

The two other groups of linked mutations, where both non-synonymous and synonymous mutations were observed, are more difficult to interpret and seem to generate both more susceptible and more resistant isolates, resulting in a skewed distribution of MIC values. Although statistically significant within our ordinal regression framework, this suggests that these latter mutations are not directly associated with a change in the MIC values but rather with a specific group of *oprD* haplotypes. In examining efflux pump genes associated to meropenem resistance, we found a frameshift mutation within *mexZ*, - “mexZ_489_−” (Figure 3) carried by 12 isolates among which 10 have a MIC value >8ug/ml. Surprisingly, this frameshift correlates with an increase in MIC which appears unexpected. In the same way, but to a weaker extend, the early stop mutation found in *mexX* (mexX:Q395*) observed in 8 isolates correlated with a slight increase in the MICs of those isolates.

Conversely, the synonymous *mexD* mutation (mexD_1752_c) observed in 6 isolates was found to be associated with a decrease in MIC. All these results in efflux pump genes must be treated with caution since the mutation frequencies are quite low. Finally, we identified a non-synonymous mutation in *ampO* - ampO:P287H (Figure3), yielding to a decrease in MIC values for the 6 isolates carrying this genotype. All associations results are listed in Supplementary Tables 2 and 3).

**Discussion**

The susceptibility of a microbial pathogen to therapeutically relevant antibiotics is one of the clinically most relevant bacterial phenotypes. Genome sequences have already been shown to confirm existing and facilitate the discovery of new resistance traits [7, 9]. Huge databases have been developed that contain more or less complete inventories of all genetic factors known to be involved in antimicrobial resistance. The format of such database may differ and some of them are accompanied by specific software packages that facilitate aimed searches for resistance genes in genome sequences or metagenomic databases [113, 114, 115, 116, 117]. Tools such as these are relatively new and still need specific testing for the various bacterial species. We here focus on *P. aeruginosa* and its MDR and XDR clones as a model bacterial species and defined the genomic polymorphisms associated with antibiotic (multi-)resistance. In particular, *oprD* was confirmed as an important meropenem AR gene. Globally, our analyses confirm the correlation between MICs and most of the known resistance traits if these occur in sufficient frequency in our strain panel. Our statistical framework made it possible to explore even more detailed associations by controlling for the already established correlations. Moreover, all association studies were strictly controlled for population structure via a principal component driven approach based on a calculation of core SNP co-variances. As such the approach presented here is fully bacterial species' independent and new mutations, including ones involved in genetic control rather than being directly causal, can be discovered. This was illustrated here by the mutations found within *oprD*, in particular the two mutations leading to an early stop codon and associated with an increase of meropenem MICs.

Intrinsic resistance is an important and extensive phenotype for *P. aeruginosa* [120]. It's genomic markers are many and diverse and gaining additional insight in the intrinsic resistance gene array is still important and clonal complexes may be helpful

in delineating full intrinsic resistomes. Recently, the first genome sequences of representatives of the *P. aeruginosa* MDR clone ST111 were reported [25]. Comparative analysis of these sequences with regard to genes involved in antibiotic or antiseptic resistance revealed key resistance mutations in a variety of well-known genes. The authors suggested that rapid mapping of the polymorphisms detected could be used for optimization of antibiotic choice. Similar data for ST235 became available and genomic islands 1 and 2 were shown to be important in the dynamics of resistance development. Class 1 integrons, already before shown to often carry metallo-beta-lactamase genes [119], were identified as important elements and integration and conjugation as important evolutionary processes facilitating resistance flexibility [121]. Even more recently a detailed study described the successful detection of presumptive ceftazidime resistance markers in 88% of all *P. aeruginosa* genomes studied [118]. Here we successfully introduce a method that does not only allow for monitoring resistance but which is even able to categorize strains on the basis of elevated MICs, even below the clinically relevant resistance level. In the context of elevated LD causing “statistical fog” we are still able to exploit the sequence information collected and identify mutations that correlate well with the MIC distribution. This allows for the identification of mutations which in themselves do not cause resistance but which, when accumulating over time, may still result in clinical resistance. This generates information on the emergence and evolution of new haplotypes that may even further accumulate mutations. It may also help understand the mechanisms of resistance development including genetic control and regulatory feedback loops. More than the validation of known mechanisms, the resistome-wide association study allowed to highlight new candidates, such as mexX/Y efflux pump for amikacin) together with the importance of the allelic form of accessory genes in correlation with MIC increase or decrease (*oprD* forms in meropenem resistance and OXA forms in cefepime resistance). This study also illustrated some limits in a single isolate panel constitution for multi-drug testing: low sample size (around 100 isolates), low prevalence of a resistance status (below 3%) and very low allele frequencies prevent from detecting good associations, due to a lack of statistical power.

Additional analyses using phenotypically well-defined and genetically diverse as well as ST-identical strains need to be performed in order to validate our initial findings. When this has been done we may then move towards real genome-wide association

studies (GWAS) without the strict need for a reference genome, which would otherwise be quite restrictive for highly plastic species such as *P. aeruginosa*. Such methods may further detail genotype-phenotype associations but care must be taken: what works well for clonal species may still be difficult for species with an a more panmictic population structure. A realistic assessment of feasibility must be done, way before thinking of introducing such methods in the routine microbiology laboratory.

**Acknowledgments**

The authors thank Chris Huff and Kathy Wilkey (bioMerieux St. Louis, USA) for their help with the cultivation and shipment of the *P. aeruginosa* strains.

**Funding**

The entire study was financed using internal funds available to the authors.

**Transparency Declarations**

M.J., A.v.B., D.C., D.S., B.B., W.M.D.Jr, G.Z., N.M., P.M., C.I.P., S.S., G.G. and J.B.V. are employees of bioMerieux (Marcy l'Etoile, France) , an international company developing and selling in vitro diagnostics for infectious diseases. K.C. and M.B. are employees of Synthetic Genomics Inc (La Jolla, USA), a company working in the field of synthetic genomics and associated therapies.

**References (check for redundancies)**

**3. Cho HH, Kwon KC, Kim S, Koo SH.** Correlation between virulence genotype and fluoroquinolone resistance in carbapenem-resistant Pseudomonas aeruginosa. Ann Lab Med. 2014 Jul;34(4):286-92. doi: 10.3343/alm.2014.34.4.286. Epub 2014 Jun 19.

**7. Dordel J, Kim C, Chung M, Pardos de la Gándara M, Holden MT, Parkhill J, de Lencastre H, Bentley SD, Tomasz A.** Novel determinants of antibiotic resistance: identification of mutated loci in highly methicillin-resistant subpopulations of methicillin-resistant Staphylococcus aureus. MBio. 2014 Apr 8;5(2):e01000. doi: 10.1128/mBio.01000-13.

**9. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, Turner P, Bentley SD, Parkhill J.** Comprehensive Identification of Single Nucleotide Polymorphisms Associated with

Beta-lactam Resistance within Pneumococcal Mosaic Genes. PLoS Genet. 2014 Aug 7;10(8):e1004547. doi: 10.1371/journal.pgen.1004547. eCollection 2014 Aug.

**10. Singh G, Srinivasan R, Cheng J, Peng Z, Fujimura K, Baek MS, Panzer AR, Tringe SG, Chen F, Sorek R, Weng L, Bristow J, Wiener-Kronish JP, Lynch SV.** Rearrangement of a large novel Pseudomonas aeruginosa gene island in strains isolated from a patient developing ventilator-associated pneumonia. J Clin Microbiol. 2014 Jul;52(7):2430-8. doi: 10.1128/JCM.01626-13. Epub 2014 Apr 30.

**13. Cabot G1, Ocampo-Sosa AA, Domínguez MA, Gago JF, Juan C, Tubau F, Rodríguez C, Moyà B, Peña C, Martínez-Martínez L, Oliver A; Spanish Network for Research in Infectious Diseases (REIPI).** Genetic markers of widespread extensively drug-resistant Pseudomonas aeruginosa high-risk clones. Antimicrob Agents Chemother. 2012 Dec;56(12):6349-57. doi: 10.1128/AAC.01388-12. Epub 2012 Oct 8.

**14. Guzvinec M, Izdebski R, Butic I, Jelic M, Abram M, Koscak I, Baraniak A, Hryniewicz W, Gniadkowski M, Tambic Andrasevic A.** Sequence Types 235, 111, and 132 Predominate among Multidrug-Resistant Pseudomonas aeruginosa Clinical Isolates in Croatia. Antimicrob Agents Chemother. 2014 Oct;58(10):6277-83. doi: 10.1128/AAC.03116-14. Epub 2014 Jul 28.

**17.Woodford N, Turton JF, Livermore DM.** Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS Microbiol Rev. 2011 Sep;35(5):736-55. doi: 10.1111/j.1574-6976.2011.00268.x. Epub 2011 Mar 1. Review.

**25. Witney AA, Gould KA, Pope CF, Bolt F, Stoker NG, Cubbon MD, Bradley CR, Fraise A, Breathnach AS, Butcher PD, Planche TD, Hinds J.** Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of Pseudomonas aeruginosa. Clin Microbiol Infect. 2014 Jan 14. doi: 10.1111/1469-0691.12528. [Epub ahead of print]

**27. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, Larsen MV, Aarestrup FM.** Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing.J Antimicrob Chemother. 2013 Apr;68(4):771-7. doi: 10.1093/jac/dks496. Epub 2012 Dec 11.

**28. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TE, Crook DW.** Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother. 2013 Oct;68(10):2234-44. doi: 10.1093/jac/dkt180. Epub 2013 May 30.

**29. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TE, Crook DW, Walker AS, Golubchik T.** Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. J Clin Microbiol. 2014 Apr;52(4):1182-91. doi: 10.1128/JCM.03117-13. Epub 2014 Feb 5.

**30. Kos VN, Déraspe M, McLaughlin RE, Whiteaker JD, Roy PH, Alm RA, Corbeil J, Gardner H.** The resistome of Pseudomonas aeruginosa in relationship to phenotypic susceptibility. Antimicrob Agents Chemother. 2014 Nov 3. pii: AAC.03954-14. [Epub ahead of print]

**33. Lister PD, Wolter DJ, Hanson ND.** Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. Clin Microbiol Rev. 2009 Oct;22(4):582-610. doi: 10.1128/CMR.00040-09. Review

**34. Ferreira ML, Dantas RC, Faria AL, Gonçalves IR, Silveira de Brito C, Queiroz LL, Gontijo-Filho PP, Ribas RM.** Molecular epidemiological survey of the quinolone- and carbapenem-resistant genotype and its association with the type III secretion system in Pseudomonas aeruginosa. J Med Microbiol. 2015 Mar;64(Pt 3):262-71. doi: 10.1099/jmm.0.000023. Epub 2015 Jan 16.

**35. Cholley P, Thouverez M, Hocquet D, van der Mee-Marquet N, Talon D, Bertrand X.** Most multidrug-resistant Pseudomonas aeruginosa isolates from hospitals in eastern France belong to a few clonal types. J Clin Microbiol. 2011 Jul;49(7):2578-83. doi: 10.1128/JCM.00102-11. Epub 2011 May 18

**38. van Belkum A, Dunne WM Jr.** Next-generation antimicrobial susceptibility testing. J Clin Microbiol. 2013 Jul;51(7):2018-24. doi: 10.1128/JCM.00313-13. Epub 2013 Mar 13. Review

**39. Naas T, Poirel L, Nordmann P.** Molecular characterisation of In51, a class 1 integron containing a novel aminoglycoside adenylyltransferase gene cassette, aadA6, in Pseudomonas aeruginosa. Biochimica et Biophysica Acta 1489 (1999) 445-451.

**40. Gillings MR.** Integrons: Past, Present, and Future. Microbiology and Molecular Biology Reviews. 2014 June; Vol 78(2) 257–277.

**41. Freschi L et al.** Clinical utilization of genomics data produced by the international Pseudomonas aeruginosa consortium. Frontiers in Microbiology; 2015 Sept. Vol 6(1036)

**42. Kung VL, Ozer EA, Hauser AR.** The accessory genome of Pseudomonas aeruginosa. Microbiol Mol Biol Rev 2010, 74:621-641.

**43. Roy, PH.** Genetic Mechanisms of Transfer of Drug Resistance, in Antimicrobial Drug Resistance: Mechanisms of Drug Resistance, Volume 1. Book edited by Douglas Mayers. 2009; Chapter 5, 53-62

**44. Igbinosa EO, Obuekwe IS**. Evaluation of Antibiotic Resistant Gene in Abattoir Environment.J. Appl. Sci. Environ. Manage. June 2014. Vol. 18 (2) 165-170

**53. Popescu AA, Harper AL, Trick M, Bancroft I, Huber KT.** A novel and fast approach for population structure inference using kernel-PCA and optimization. Genetics. 2014 Dec;198(4):1421-31. doi: 10.1534/genetics.114.171314

**45.** McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society, Series B 42, 109–142.

**46.** Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing" (PDF). Journal of the Royal Statistical Society, Series B 57 (1): 289–300. MR 1325392

**99. van Belkum A, Soriaga LB, LaFave MC, Akella S, Veyrieras JB, Barbu EM, Shortridge D, Blanc B, Hannum G, Zambardi G, Miller K, Enright MC, Mugnier N, Brami D, Schicklin S, Felderman M, Schwartz AS, Richardson TH, Peterson TC, Hubby B, Cady KC.** Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant Pseudomonas aeruginosa. MBio. 2015 Nov 24;6(6):e01796-15.

100. Katoh, Misawa, Kuma, Miyata 2002 (Nucleic Acids Res. 30:3059-3066) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.

**101. Akaike, H.** (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B.N.; Csáki, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, p. 267-281.

**102. Wilks, S. S.** (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". The Annals of Mathematical Statistics 9: 60–62

**103. Turton JF, Wright L, Underwood A, Witney AA, Chan YT, Al-Shahib A, Arnold C, Doumith M, Patel B, Planche TD, Green J, Holliman R, Woodford N.** High-Resolution Analysis by Whole-Genome Sequencing of an International Lineage (Sequence Type 111) of Pseudomonas aeruginosa Associated with Metallo-Carbapenemases in the United Kingdom. J Clin Microbiol. 2015 Aug;53(8):2622-31.

**104. Wiehlmann L1, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Köhler T, van Delden C, Weinel C, Slickers P, Tümmler B.** Population structure of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A. 2007 May 8;104(19):8101-6.

**105. Witney AA, Gould KA, Pope CF, Bolt F, Stoker NG, Cubbon MD, Bradley CR, Fraise A, Breathnach AS, Butcher PD, Planche TD, Hinds J.** Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of Pseudomonas aeruginosa. Clin Microbiol Infect. 2014 Oct;20(10):O609-18.

**106. van Belkum A, Dunne WM Jr.** Next-generation antimicrobial susceptibility testing. J Clin Microbiol. 2013 Jul;51(7):2018-24.

**107. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CL, Bowden R, Drobniewski FA, Allix-Béguec C, Gaudin C, Parkhill J, Diel R, Supply P, Crook DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TE; Modernizing Medical Microbiology (MMM) Informatics Group.** Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015 Oct;15(10):1193-202.

**108. Zhao S, Tyson GH, Chen Y, Li C, Mukherjee S, Young S, Lam C, Folster JP, Whichard JM, McDermott PF.** Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in Campylobacter spp. Appl Environ Microbiol. 2015 Oct 30;82(2):459-66.

**109. European Committee on Antimicrobial Susceptibility Testing.** Consultation on Report from the EUCAST subcommittee on the role of whole genome sequencing (WGS) in antimicrobial susceptibility testing of bacteria. May 2016.

**110. Pirnay JP, Bilocq F, Pot B, Cornelis P, Zizi M, Van Eldere J, Deschaght P, Vaneechoutte M, Jennes S, Pitt T, De Vos D.** Pseudomonas aeruginosa population structure revisited. PLoS One. 2009 Nov 13;4(11):e7740.

**111. Oliver A, Mulet X, López-Causapé C, Juan C.** The increasing threat of Pseudomonas aeruginosa high-risk clones. Drug Resist Updat. 2015 Jul-Aug;21-22:41-59.

**112. Aubert, D., L. Poirel, J. Chevalier, S. Leotard, J. M. Pages, and P. Nordmann. 2001.** Oxacillinase-mediated resistance to cefepime and susceptibility to ceftazidime in Pseudomonas aeruginosa. Antimicrob. Agents Chemother. 45:1615-1620.

**113. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD.** The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013 Jul;57(7):3348-57.

**114. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, de Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismail N, Omar SV, Smith EG, Buck D, McVean G, Walker AS, Peto TE, Crook DW, Iqbal Z.** Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun. 2015 Dec 21;6:10063.

**115. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012 Nov;67(11):2640-4.

**116. Xavier BB, Das AJ, Cochrane G, De Ganck S, Kumar-Singh S, Aarestrup FM, Goossens H, Malhotra-Kumar S.** Consolidating and Exploring Antibiotic Resistance Gene Data Resources. J Clin Microbiol. 2016 Apr;54(4):851-9.

**117. Rowe W, Baker KS, Verner-Jeffreys D, Baker-Austin C, Ryan JJ, Maskell D, Pearce G.** Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. PLoS One. 2015 Jul 21;10(7):e0133492.

**118. Kos VN, McLaughlin RE, Gardner HA.** Elucidation of Mechanisms of Ceftazidime Resistance among Clinical Isolates of Pseudomonas aeruginosa by Using Genomic Data. Antimicrob Agents Chemother. 2016 May 23;60(6):3856-61.

**119. Wright LL, Turton JF, Hopkins KL, Livermore DM, Woodford N.** Genetic environment of metallo-β-lactamase genes in Pseudomonas aeruginosa isolates from the UK. J Antimicrob Chemother. 2015 Dec;70(12):3250-8.

**120. Murray JL, Kwon T, Marcotte EM, Whiteley M.** Intrinsic Antimicrobial Resistance Determinants in the Superbug Pseudomonas aeruginosa. MBio. 2015 Oct 27;6(6):e01603-15.

**121. Roy Chowdhury P, Scott M, Worden P, Huntington P, Hudson B, Karagiannis T, Charles IG, Djordjevic SP.** Genomic islands 1 and 2 play key roles in the evolution of extensively drug-resistant ST235 isolates of Pseudomonas aeruginosa. Open Biol. 2016 Mar;6(3). pii: 150175. doi: 10.1098/rsob.150175.

## FIGURE LEGENDS

**Figure 1. Overview of the phenotypic, genomic and resistome data of the 672 *P. aeruginosa* isolates included in the present analysis.** The phylogenetic tree is inferred from core-gene content and depicts the 3 major groups of *P. aeruginosa* in shades of purple. Depicted below the tree in dark green bars, are the isolates which were found resistant to one or more of the 9 antibiotics analyzed in this study. Below this part of the graph, strains with genomes larger or smaller than the median are depicted as black or grey, respectively. The left panel provides the names of genes and non-synonymous mutations constituting the resistome for the 9 key antibiotics. Different antibiotic resistance mechanisms are color coded (bar to the right of the gene and mutation list). The right box reviews gene content on a per isolate basis and endogenous resistance markers can be easily distinguished from the acquired ones by gray shading. The resistome structure defines core (45%) versus accessory (55%) resistance elements, where it is obvious that antibiotic inactivation genes are more likely to belong to the accessory resistome than efflux genes. Note that resistance genes embedded in integrons are color coded as dark red bars. The percentages on the far right define gene prevalence across the entire group of strains.

**Figure 2. Class-I integron composition analysis for *P. aeruginosa***. Since only IntI1 sequences were detected in our panel, only class-I integrons are reported. Panel A depicts the integron frequency of occurrence (34.1% in our case) in the isolates analyzed and assesses the AR content of putative integrons by searching for resistance genes up to 10kb upstream of the IntI1 motif. The resistance genes were tested for enrichment in integrons compared to alternative genome locations. Panel B further identify the order of the genes and their individual frequency of occurrence for the main integron cassette, In51. Panel C depicts the different configurations (in terms of AR gene composition) of the most frequent IntI1 integrons, with the most frequent one (n=27) containing sul1 and aadA6 AR genes only. Panel D shows their presence in the different clonal *P. aeruginosa* clusters as present in the phylogenetic tree. Note that some integron types cluster according to bacterial sT, whereas others are more widely scattered throughout the phylogenetic tree.

**Figure 3.** – **Synoptic representation of the Resistome Wide Association Study (RWAS) results**. Only genes and mutations either known to be associated or subsequently found to be associated with resistance against the 9 antibiotics tested here are displayed. Panel A provides the results for the presence versus absence of genes while panel B focuses on specific mutations within genes. Genes and mutations are reported on columns and antibiotics in rows. Note that rows were ordered by decreasing phenotype prevalence, and in panel A, columns were ordered by increasing allele frequency. In both panels, an empty cell in the table means that no association was found. Green cell indicates that a significant new association has been detected. Dark or light red cell stands for a known determinant found to be significantly associated with MIC variability. Only determinants whose cells are filled with dark red have been retained by our backward-elimination procedure to test for new associations. Conversely, dark or light blue cell stands for known determinants, but these are not significantly associated with the MIC variability in our panel of strains. Nevertheless some of these determinants (whose cells are filled with dark blue) were retained by our conservative backward-elimination procedure (see Material and Methods). Finally, the nature of the mutation is color coded on the top of panel B. The bottom panel represents the extent of linkage disequilibrium (LD) between mutations.