

1 **Insights from the revised complete genome sequences of** 2 ***Acinetobacter baumannii* strains AB307-0294 and ACICU** 3 **belonging to global clone 1 and 2** 4

5 **1.1 Author names**

6 Mohammad Hamidian*¹, Ryan Wick², Rebecca M. Hartstein³, Louise Judd², Kathryn E.
7 Holt^{2,4} and Ruth M. Hall³

8 **1.2 Affiliation**

9 ¹The ithree institute, University of Technology Sydney, Ultimo, NSW, Australia;

10 ²Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne,
11 Victoria 3004, Australia; ³School of Life and Environmental Sciences, The University of
12 Sydney, Australia; ⁴London School of Hygiene & Tropical Medicine, London WC1E 7HT,
13 UK.

14 **1.3 Corresponding author**

15 * mohammad.hamidian@uts.edu.au

16 **1.4 Keyword**

17 *Acinetobacter baumannii*, AB307-0294, ACICU, global clone 1, GC1, global clone 2, GC2,
18 complete genome sequence and Whole Genome Shotgun (WGS).

19 **1.5 Repositories:**

20 The complete genome sequences have been deposited in DDBJ/ENA/GenBank under the
21 GenBank accession numbers CP001172.2 (AB307-0294 chromosome), CP031380 (ACICU
22 chromosome), CP031381 (pACICU1) and CP031382 (pACICU2).

23 **2. Abstract**

24 The *Acinetobacter baumannii* global clone 1 (GC1) isolate AB307-0294, recovered in the
25 USA in 1994, and the global clone 2 (GC2) isolate ACICU, isolated in 2005 in Italy, were
26 among the first *A. baumannii* isolates to be completely sequenced. AB307-0294 is susceptible
27 to most antibiotics and has been used in many genetic studies and ACICU belongs to a rare
28 GC2 lineage. The complete genome sequences, originally determined using 454
29 pyrosequencing technology which is known to generate sequencing errors, were re-
30 determined using Illumina MiSeq and MinION (ONT) technologies and a hybrid assembly
31 generated using Unicycler. Comparison of the resulting new high-quality genomes to the
32 earlier 454-sequenced version identified a large number of nucleotide differences affecting
33 protein coding features, and allowed the sequence of the long and highly-repetitive *bap* and
34 *blp1* genes to be properly resolved for the first time in ACICU. Comparisons of the
35 annotations of the original and revised genomes revealed a large number of differences in the
36 protein coding features (CDSs), underlining the impact of sequence errors on protein
37 sequence predictions and core gene determination. On average, 400 predicted CDSs were
38 longer or shorter in the revised genomes and about 200 CDS features were no longer present.

39 **3. Impact statement**

40 The genomes of the first 10 *A. baumannii* strains to be completely sequenced underpin a large
41 amount of published genetic and genomic analysis. However, most of their genome
42 sequences contain substantial numbers of errors as they were sequenced using 454
43 pyrosequencing, which is known to generate errors particularly in homopolymer regions; and
44 employed manual PCR and capillary sequencing steps to bridge contig gaps and repetitive
45 regions in order to finish the genomes. Assembly of the very large and internally repetitive
46 gene for the biofilm-associated proteins Bap and BLP1 was a recurring problem. As these

47 strains continue to be used for genetic studies and their genomes continue to be used as
48 references in phylogenomics studies including core gene determination, there is value in
49 improving the quality of their genome sequences. To this end, we re-sequenced two such
50 strains that belong to the two major globally distributed clones of *A. baumannii*, using a
51 combination of highly-accurate short-read and gap-spanning long-read technologies.
52 Annotation of the revised genome sequences eliminated hundreds of incorrect CDS feature
53 annotations and corrected hundreds more. Given that these revisions affected hundreds of
54 non-existent or incorrect CDS features currently cluttering GenBank protein databases, it can
55 be envisaged that similar revision of other early bacterial genomes that were sequenced using
56 error-prone technologies will affect thousands of CDS currently listed in GenBank and other
57 databases. These corrections will impact the quality of predicted protein sequence data stored
58 in public databases. The revised genomes will also improve the accuracy of future genetic
59 and comparative genomic analyses incorporating these clinically important strains.

60 **4. Data summary**

- 61 1. The corrected complete genome sequence of *A. baumannii* AB307-0294 has been
62 deposited in GenBank; GenBank accession number CP001172.2 (chromosome url -
63 <https://www.ncbi.nlm.nih.gov/nuccore/CP001172.2>).
 - 64 2. The corrected complete genome sequence of ACICU has been deposited in GenBank
65 under the GenBank accession numbers CP031380 (chromosome; url -
66 <https://www.ncbi.nlm.nih.gov/nuccore/CP031380>), CP031381 (pACICU1; url -
67 <https://www.ncbi.nlm.nih.gov/nuccore/CP031381>) and CP031382 (pACICU2; url -
68 <https://www.ncbi.nlm.nih.gov/nuccore/CP031382>).
- 69 **The authors confirm all supporting data, code and protocols have been provided within**
70 **the article or through supplementary data files.**

71 **5. Introduction**

72 *Acinetobacter baumannii* is a Gram-negative bacterium that has emerged as an important
73 opportunistic pathogen and is a research priority because of its high levels of resistance to
74 antibiotics (1-3), desiccation, and heavy metals (4, 5). On a global scale, members of two
75 clinically important clones, known as global clone 1 (GC1) and global clone 2 (GC2), have
76 been responsible for the majority of outbreaks caused by multiply antibiotic resistant *A.*
77 *baumannii* strains (1-3, 6-8). Whole genome sequencing (WGS) technologies have
78 revolutionised the study of bacterial pathogens allowing the entire gene repertoire of bacterial
79 strains to be determined and hence enabling the study of the relationships between outbreak
80 strains with an unprecedented high resolution (9). However, accuracy is important.

81 The first 10 complete genomes of *A. baumannii* strains were reported between 2006-
82 2012 (Table 1) and are still used as baseline in many studies of this microorganism (10-12).
83 Except for three strains (AYE, TCDC-AB0715 and TYTH-1), all of the early *A. baumannii*
84 complete genomes were sequenced using the 454-pyrosequencing technology and assembled
85 using PCR. Pyrosequencing is known to generate frequent systematic sequencing errors,
86 especially errors in the length of homopolymeric runs (13); and these errors lead to erroneous
87 protein sequence (CDS) prediction, often associated with fragmentation of genuine open
88 reading frames.

89 An additional problem in *A. baumannii* genomes determined using short read
90 sequence data followed by PCR gap closure arises from the many short internal repeats
91 present in the very large *bap* gene (~8-25 kbp), which is hard to assemble accurately. This
92 gene encodes the biofilm associated protein Bap (14-17). The *bap* gene was originally cloned
93 from AB307-0294 (GC1), and found to be 25,863 bp with a complex configuration of
94 internal repeats (15). However, the size of the *bap* gene from a GC2 isolate was estimated at
95 approximately 16 kbp (16). In another study, the length of Bap proteins predicted from *A.*

96 *baumannii* genomes available in GenBank appeared to be highly variable, mainly due to
97 different numbers of copies of the various repeated segments and the reading frame was often
98 fragmented (17). The *blp1* gene, which is 9-10 kbp encodes a further very large protein that
99 also has internal repeats and is associated with biofilm formation (17).

100 Newer sequencing technologies such as PacBio (Pacific Biosciences) and MinION
101 (Oxford Nanopore Technologies, ONT) can generate much longer sequencing reads (9)
102 allowing gaps to be spanned. MinION only assemblies are also prone to errors (18) but can
103 be combined with high-accuracy Illumina short read data to produce very high quality
104 finished genome assemblies (19). Long read sequence data has enabled a re-assessment of
105 early completed *A. baumannii* genomes, including several of the first 10 to be sequenced
106 (Table 1). For example, in 2016, ATCC 17978 was re-sequenced using PacBio. This revealed
107 the presence of a 148 kb conjugative plasmid, pAB3, fragments of which were erroneously
108 merged into the chromosome in the original 454-based assembly (20). This plasmid sequence
109 brought together the parts of *GI_{sul2}*, fragmented pieces of which had been randomly
110 distributed in the chromosome in the original sequence (21). In 2017, we revised the 454-
111 based genome sequence of the GC1 strain AB0057 using Illumina HiSeq technology, and
112 found that hundreds of single base additions or deletions changed >200 protein coding
113 features (CDSs) (22). An additional copy of the *oxa23* carbapenem resistance gene, located in
114 Tn2006, was also found in the revised sequence of the chromosome (GenBank no.
115 CP001182.2) (22, 23).

116 A recent revision of the 454-based genome of the GC2 strain MDR-ZJ06 using
117 PacBio sequencing led to the correction of hundreds of CDS features and allowed
118 reassessment of the localisation of important antimicrobial resistance regions (24). The
119 position of transposon Tn2009, which carries the *oxa23* gene, was revised; and a region
120 originally reported as a plasmid, but that had been predicted to be a chromosomally-located

121 AbGRI3 type resistance island (25), was incorporated into the chromosome (CP001937.2)
122 (24). In the revised genome, the two arrays of gene cassettes carrying antibiotic resistance
123 genes in class 1 integrons are now in the correct resistance islands. These revisions exemplify
124 the challenges encountered when relying solely on short read data to assemble bacterial
125 genomes and highlight the extent and impact of pyrosequencing errors particularly on CDS
126 predictions.

127 Two further *A. baumannii* strains for which only early 454-based genome sequences
128 are available are the largely antibiotic susceptible isolate AB307-0294, recovered from the
129 blood of a patient hospitalized in Buffalo, NY, in 1994 (26), and the extensively antibiotic
130 resistant isolate ACICU recovered in 2005 from cerebrospinal fluid of patient in San
131 Giovanni Addolorata Hospital in Rome, Italy (GenBank no. CP000863) (27). AB307-0294
132 was one of the first global clone 1 (GC1) strains to be completely sequenced (26) and has
133 been extensively used in genetic studies (28-32). It belongs to CC1 (ST1) in the Institut
134 Pasteur multi-locus sequence typing (MLST) scheme and to ST231 in the Oxford MLST
135 scheme and carries the KL1 capsule genes and OCL1 at the outer core locus (33) (Table 1).
136 Compared to other GC1 strains characterised to date, AB307-0294 is relatively susceptible to
137 antibiotics (26), exhibiting resistance only to chloramphenicol (intrinsic) and nalidixic acid
138 (acquired). It contains no plasmids.

139 ACICU was the first global clone 2 (GC2) isolate to be sequenced (27). It belongs to
140 ST2 in the Institut Pasteur MLST scheme and carries the KL2 capsule genes and OCL1 at
141 the outer core locus (34). ACICU is carbapenem resistant and also resistant to multiple
142 antibiotics including third generation cephalosporins, sulfonamides, tetracycline, amikacin,
143 kanamycin, netilmicin and ciprofloxacin (27). It contains two plasmids (27). However, we
144 previously showed that the largest plasmid, pACICU-2, which was reported to include no
145 resistance genes, is larger and contains the amikacin resistance gene *aphA6* in transposon

146 *TnaphA6*. The central segment of *TnaphA6*, including the *aphA6* gene and one of the
147 ISAbal25 copies as well as a 4.7 kb backbone segment were missing in the original 454-
148 based whole genome sequence (35).

149 Here, we report revised complete genome sequences for *A. baumannii* strains AB307-
150 0294 (GC1) and ACICU (GC2), generated using MiSeq (Illumina) and MinION (ONT)
151 sequence data. The new genome sequences correct hundreds of protein coding features
152 generated by the presence of SNDs (single nucleotide differences) and small
153 insertion/deletions of mainly 1-3 bases in the earlier 454 genome sequences.

154 **6. Methods**

155 **6.1 Whole genome sequencing, assembly and annotation**

156 Whole cell DNA was isolated and purified using the protocol described previously (1, 36).
157 Libraries were prepared from whole cell DNA isolated from AB307-0294 and ACICU and
158 were sequenced using Illumina MiSeq and ONT MinION. Paired-end reads of 150 bp and
159 MinION reads of up to 20 kb were used to assemble each genome using the Unicycler
160 software (v0.4.0) (19) using default parameters.

161 Protein coding, rRNA and tRNA genes were annotated using the automatic annotation
162 program Prokka v1.13 (37). Regions containing antibiotic resistance genes and the
163 polysaccharide biosynthesis loci, biofilm-associated proteins and genes used in the MLST
164 schemes were annotated manually.

165 To compare previous CDS (≥ 25 aa CDS features) annotations with our new results, we wrote
166 a script (github.com/rrwick/Compare-annotations) to quantify the differences. This script
167 classifies coding sequences in the annotations as either exact matches, inexact matches, only
168 present in the first annotation or only present in the second annotation. We also used the Ideel
169 pipeline of Dr Mick Watson (github.com/mw55309/ideel) to assess the completeness of CDS

170 annotated in each genome, by comparing the length of each CDS to that of its longest
171 BLAST hit in the Uniprot database (as described in [http://www.opiniomics.org/a-simple-test-
172 for-uncorrected-insertions-and-deletions-indels-in-bacterial-genomes/](http://www.opiniomics.org/a-simple-test-for-uncorrected-insertions-and-deletions-indels-in-bacterial-genomes/)).

173

174 **7. Results and discussion**

175 **7.1 Revised genome of ACICU**

176 ACICU, the first GC2 strain to be completely sequenced, contains AbaR2 in the
177 chromosomal *comM* gene (27). As this AbaR resistance island type is more usually found in
178 this location in GC1 strains (38) with an AbGR11 type island in GC2 isolates (39), ACICU
179 may represent a rare GC2 lineage. Here, the ACICU genome was re-sequenced using a
180 combination of Illumina (MiSeq, 58x depth) and ONT (MinION, 253x depth) data. The new
181 contiguous ACICU chromosomal sequence comprised 3,919,274 bp (GenBank no.
182 CP001172.2), compared to 3,904,116 bp in the original submission (GenBank no. CP000863),
183 making the revised chromosome 15,158 bp longer (Table 1). Most of the additional length in
184 the revised chromosome was found to be due to a 11.2 kbp longer *bap* gene, which is just over
185 11 kbp and in 9 smaller orfs in the original sequence (locus_ids ACICU_02938 to
186 ACICU_2946) as noted previously (17). In the revised genome sequence the *bap* gene is 22.2
187 kbp (BAP; locus_id DMO12_08904), mainly due to a large number of short strings of repeated
188 sequences missing previously. Hence, some of the variation in length of *bap* reported
189 previously (17) may be due to sequencing and assembly issues rather than genuine length
190 variation in the *A. baumannii* population. The *blp1* gene in the original sequence (locus_id
191 ACICU_02910) is 9510 bp and 9813 bp (locus_id DMO12_08811) in the revised genome.

192 The revised chromosome of ACICU differs from the original at 281 positions including
193 40 SNDs and 241 insertions or deletions of 1-3 bases (mostly in homopolymeric runs of As or

194 Ts). The original annotation included 3677 protein-coding features (CDS features are ≥ 25 aa)
195 whereas the revised genome annotation contains 3605 CDS features. Comparison of the CDS
196 features indicated that only 3129 CDSs are identical between the two versions. The differences
197 are mostly due to correction of open reading frames that were interrupted or fused due to errors
198 in the 454 sequence and include 80 CDSs unique to the revised version and 142 CDS features
199 in the original sequence that could not be found in the corrected chromosome. A further 396
200 CDS that are present in both versions are altered: of these, 8 have the same length, 285 are
201 longer in the revised chromosome and 103 are shorter. Overall, 98.8% of all genes (n=3568)
202 in the new assembly are within 5% of the maximum length of homologous proteins in Uniprot
203 (i.e. the expected length), calculated using the ideel pipeline (see Methods). In the old
204 assembly, only 95.8% (n=3494) of all genes are within 5% of this expected length. The
205 distribution of length ratios is shown in Fig. 1A, highlighting a substantial population of CDS
206 annotated in the old assembly that have lengths well below those of homologous proteins in
207 Uniprot.

208 ACICU carries two plasmids (Table 1), pACICU1 and pACICU2 (27), which encode
209 the RepAci1 and RepAci6 replication initiation proteins (40). The original pACICU1 sequence
210 (GenBank no. CP000864) is 28279 bp long and contains two copies of the carbapenem
211 resistance gene *oxa58* while the revised pACICU1 (GenBank no. CP031381) is 24268 bp long
212 and includes only a single *oxa58* copy. It lacks the region between the two IS26 and one copy
213 of IS26 in the original sequence. The IS26 mediated duplication may have been generated
214 during growth in selective media. The original and revised pACICU1 sequences also differed
215 by 3 SNDS, 6 single bp insertions and 1 single bp and 2 of 2 bp deletions. We previously used
216 a PCR mapping strategy (35) to show that the *aphA6* gene and an additional ISAbal25 as well
217 as a 4.7 kb long backbone segment, located between two copies of a ~420 bp repeated segment,
218 are missing from the original sequence of pACICU2, the larger plasmid of ACICU (35). Here,

219 the long-read sequences generated for pACICU2 (GenBank no. CP031382) confirmed this.
220 The revised plasmid sequence differs by 6 SND from pAb-G7-2 (GenBank no. KF669606.1),
221 a conjugative plasmid from a GC1 isolated in Australia in 2003 reported previously (41).

222 7.2 Revised genome of AB307-0294

223 The AB307-0294 genome was also sequenced using a combination of Illumina (MiSeq, 63x
224 depth) and ONT (MinION, 120x depth) technologies. The hybrid assembly resulted in a
225 single 3,759,495 bp chromosome (GenBank no. CP001172.2) compared with 3,760,981 bp in
226 the original genome (GenBank no. CP001172.1), making the revised genome 1486 bp shorter
227 (Table 1). As with AB0057, the majority of differences were found to be additions or
228 deletions of 1-3 bases, usually in “A” or “T” in homopolymeric runs of these nucleotides.
229 The original annotation included 3427 CDS while the revised annotation contains 3458 (≥ 25
230 aa), of which 2937 CDSs are identical in the two versions. Corrections of insertion/deletion
231 errors changed 354 reading frames leading to merging and splitting of CDS regions. Amongst
232 these 354 CDS features, 286 CDSs in the revised genome are longer and 65 are shorter than
233 the corresponding CDSs in the original annotation and 3 have the same length but differ
234 internally. The revised genome also includes 136 novel CDS features, compared to the
235 original sequence, while there are also 167 CDS in the old sequence that no longer exist in
236 the revised genome again indicating the high impact of the errors caused by the use of 454-
237 pyrosequencing technology. Overall, 98.9% of all genes ($n=3387$) in the new assembly are
238 within 5% of the expected length, calculated using the ideel pipeline, versus just 96.4%
239 ($n=3336$) in the old assembly (Fig. 1B).

240 The *bap* gene was 25863 bp (locus_id ABBFA_00771), the same length as reported
241 originally (15) but 1067 bp shorter than the 26930 bp *bap* gene in the original genome
242 sequence where it is split into two open reading frames (locus_id ABBFA_000776) and
243 (locus_id ABBFA_000777). The revised genome was found to contain a 10089 bp *blp1* gene

244 (ABBFA_00802), only 18 bp longer than that in the original sequence. Interestingly, both the
245 original and revised genomes appear to be devoid of any insertion sequences (IS).

246 **7.3 Revised genomes affect many predicted protein sequences**

247 To date, 6 early *A. baumannii* genome sequences, including AB307-0294 and ACICU
248 reported here, have been corrected and in each case the revised genome has resulted in
249 correction of ~ 600 CDS features on average (20, 22). In each comparison of revised and
250 original genome sequences, 100-150 new CDS features appeared, 150-200 CDSs disappeared
251 and 150-200 CDSs changed. As the extent of errors had not been reported previously (20), we
252 also compared the original (GenBank no. CP000521.1) and revised (GenBank no.
253 CP012004.1) genomes of *A. baumannii* ATCC 17978. This revealed that the revised sequence
254 has extensively re-ordered parts of the chromosome correcting a large number of inversions,
255 insertion/deletions and other mis-assemblies. A striking difference between the two genomes
256 is the inclusion in the original chromosome assembly of several large segments that in fact
257 make up a 148 kb plasmid (pAB3) carrying the *sul2* sulfonamide resistance gene (GenBank
258 no. CP012005). The misassembly issues precluded a simple alignment of the two chromosome
259 sequences, but alignment of 14 separate chromosomal segments totalling 3843892 bp, revealed
260 334 SNPs as well as 635 deletions and 754 insertions of 1-3 bases, mainly “A”s or “T”s in runs
261 of “A”s or “T”s. Overall, 3503 genes (98.2% of all genes) in the new assembly are within 5%
262 of the expected length, calculated using the ideal pipeline, versus 3381 (86.4%) in the old
263 assembly (see Fig. 1C). Hence, the original assembly was substantially flawed and should not
264 be used in future. However, although the original study reported that ATCC 17978 contains
265 two cryptic plasmids of 13 kb, pAB1 (GenBank no. CP000522.1) and 11 kb, pAB2 (GenBank
266 no. CP000523.1) (42), the revised genome does not include either of these plasmids. This may
267 be due to an assembly parameter setting to filter out the small contigs, which would remove
268 pAB1 and pAB2, from the final assembly.

269 Granted the large effects observed on the length of *bap* and *blp* in ACICU using long
270 read data, their sizes in original and revised genomes in the remainder of the first set of 10
271 (Table 1) were compared and significant differences were observed only where long read data
272 was used in the revision. In the GC2 strain MDR-ZJ06 (GenBank accession no. CP001937),
273 *blp1* (locus tag ABZJ_03096) is 9,812 bp in the revised genome (CP001937.2) versus 9,134
274 bp in the original sequence (locus tag ABZJ_03096). Further, *bap*, which is 7946 bp in the
275 revised genome (locus_id ABZJ_03955) was split into 3 orfs, ranging in size from 2 to 2.5 kb,
276 in the original sequence. In ATCC 17978, the *blp1* gene is not present in either the original or
277 the revised genome. However, the *bap* gene, which was split into two open reading frames
278 (locus_id A1S_2696; 6306 bp and A1S_2724; 1161 bp) and separated by 41 kbp in the original
279 sequence is now in a single orf (locus_id ACX60_04030; 6225 bp) in the revised genome and
280 842 bp shorter compared to those in the original genome.

281 **8. Conclusions**

282 The revised genome sequences of AB307-0294 and ACICU will underpin more accurate
283 studies of the genetics and genomic evolution of related *A. baumannii* strains belonging to
284 GC1 and GC2.

285 This work highlights the need to review and revise early bacterial genomes sequenced using
286 short read data and assembled with (or sometimes without) PCR to join contigs. Special
287 attention needs to focus on the genomes determined using the 454-pyrosequencing
288 technology in order to correct predicted protein sequences.

289 Long read data, such as those generated by PacBio and ONT (MinION) technologies, allows
290 for complete genome assembly without manual intervention. While assembling long read
291 data alone can result in sequence errors and failure to detect small plasmids, hybrid assembly
292 (using both short and long reads) can produce assemblies that are both complete and accurate.

293 However, repetitive sequences in the genome, such as the genes encoding Bap and BLP1, are
294 difficult to perfect even with hybrid assembly, so variations in these regions should be
295 interpreted with caution.

296 Finally, as the original GenBank entries are replaced by revised genomes, there is a need
297 to eliminate non-existent and incorrect predicted protein sequences in order to simplify the
298 already complex task of protein sequence searches. It can be assumed that this problem is not
299 only limited to *A. baumannii* genomes as many bacterial species so far have been sequenced
300 using the 454-pyrosequencing technology.

301 **9. Author statements**

302 **9.1 Authors and contributors**

303 Conceptualization, RMH, MH; Data curation, MH, RW; Formal analysis, MH, RW, KEH,
304 RMH; Funding, RMH, KEH, MH; Investigation, MH, RW, LJ; Resources, KEH;
305 Visualization, MH, RW, KEH; Manuscript preparation, original draft, MH and RMH; review
306 and editing RMH, MH, RW, KEH.

307 **9.2 Conflicts of interest**

308 The authors declare that there are no conflicts of interest.

309 **9.3 Funding information**

310 This work was supported by NHMRC grant GNT1079616 to RMH. MH is supported by the
311 Chancellor's Postdoctoral Research Fellowship (CPDRF PRO17-4005) from the University
312 of Technology Sydney, Australia. KEH is supported by a Senior Medical Research
313 Fellowship from the Viertel Foundation of Australia.

314 **9.4 Consent for publication**

315 Not applicable.

316 **9.5 Ethical Approval**

317 No human or animal experimentation is reported.

318 **9.6 Acknowledgements**

319 We would like to thank Prof. Thomas A. Russo, State University of New York, Buffalo, New
320 York, USA, for kindly providing AB307-0294 and Prof. Alessandra Carratoli, Istituto
321 Superiore di Sanità, Rome, Italy, for supplying ACICU.

322 **10. References**

- 323 1. Holt K, Kenyon JJ, Hamidian M, Schultz MB, Pickard DJ, Dougan G, *et al.* Five
324 decades of genome evolution in the globally distributed, extensively antibiotic-resistant
325 *Acinetobacter baumannii* global clone 1. *Microb Genom* 2016;2:e000052.
- 326 2. Post V, Hall RM. AbaR5, a large multiple-antibiotic resistance region found in
327 *Acinetobacter baumannii*. *Antimicrob Agents Chemother* 2009;53:2667-71.
- 328 3. Post V, White PA, Hall RM. Evolution of AbaR-type genomic resistance islands in
329 multiply antibiotic-resistant *Acinetobacter baumannii*. *J Antimicrob Chemother*
330 2010;65:1162-70.
- 331 4. Eijkelkamp BA, Hassan KA, Paulsen IT, Brown MH. Investigation of the human
332 pathogen *Acinetobacter baumannii* under iron limiting conditions. *BMC Genom* 2011;12:126.
- 333 5. Giannouli M, Antunes LC, Marchetti V, Triassi M, Visca P, Zarrilli R. Virulence-
334 related traits of epidemic *Acinetobacter baumannii* strains belonging to the international
335 clonal lineages I-III and to the emerging genotypes ST25 and ST78. *BMC Infect Dis*
336 2013;13:282.

- 337 6. Adams MD, Chan ER, Molyneaux ND, Bonomo RA. Genomewide analysis of
338 divergence of antibiotic resistance determinants in closely related isolates of *Acinetobacter*
339 *baumannii*. *Antimicrob Agents Chemother* 2010;54:3569-77.
- 340 7. Zarrilli R, Pournaras S, Giannouli M, Tsakris A. Global evolution of multidrug-
341 resistant *Acinetobacter baumannii* clonal lineages. *Int J Antimicrob Agents* 2013;41:11-9.
- 342 8. Wright MS, Haft DH, Harkins DM, Perez F, Hujer KM, Bajaksouzian S, *et al*. New
343 insights into dissemination and variation of the health care-associated pathogen *Acinetobacter*
344 *baumannii* from genomic analysis. *mBio* 2014;5:e00963-13.
- 345 9. Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, van Schaik W,
346 *et al*. Whole-Genome Sequencing of bacterial pathogens: the future of nosocomial outbreak
347 analysis. *Clin Microbiol Rev* 2017;30:1015-63.
- 348 10. Sahl JW, Gillece JD, Schupp JM, Waddell VG, Driebe EM, Engelthaler DM, *et al*.
349 Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to
350 pathogenesis in *Acinetobacter*. *PLoS One* 2013;8:e54287.
- 351 11. Sahl JW, Johnson JK, Harris AD, Phillippy AM, Hsiao WW, Thom KA, *et al*.
352 Genomic comparison of multi-drug resistant invasive and colonizing *Acinetobacter*
353 *baumannii* isolated from diverse human body sites reveals genomic plasticity. *BMC Genom*
354 2011;12:291.
- 355 12. Chan AP, Sutton G, DePew J, Krishnakumar R, Choi Y, Huang XZ, *et al*. A novel
356 method of consensus pan-chromosome assembly and large-scale comparative analysis reveal
357 the highly flexible pan-genome of *Acinetobacter baumannii*. *Genome Biol* 2015;16:143.
- 358 13. Balzer S, Malde K, Jonassen I. Systematic exploration of error sources in
359 pyrosequencing flowgram data. *Bioinformatics* 2011;27:i304-9.
- 360 14. Brossard KA, Campagnari AA. The *Acinetobacter baumannii* biofilm-associated
361 protein plays a role in adherence to human epithelial cells. *Infect Immun* 2012;80:228-33.

- 362 15. Loehfelm TW, Luke NR, Campagnari AA. Identification and characterization of an
363 *Acinetobacter baumannii* biofilm-associated protein. *J Bacteriol* 2008;190:1036-44.
- 364 16. Goh HM, Beatson SA, Totsika M, Moriel DG, Phan MD, Szubert J, *et al.* Molecular
365 analysis of the *Acinetobacter baumannii* biofilm-associated protein. *Appl Environ Microbiol*
366 2013;79:6535-43.
- 367 17. De Gregorio E, Del Franco M, Martinucci M, Roscetto E, Zarrilli R, Di Nocera PP.
368 Biofilm-associated proteins: news from *Acinetobacter*. *BMC Genom* 2015;16:933.
- 369 18. Watson M, Warr A. Errors in long-read assemblies can critically affect protein
370 prediction. *Nat Biotech* 2019;37:124-6.
- 371 19. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
372 assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
- 373 20. Weber BS, Ly PM, Irwin JN, Pukatzki S, Feldman MF. A multidrug resistance
374 plasmid contains the molecular switch for type VI secretion in *Acinetobacter baumannii*.
375 *Proc Natl Acad Sci U S A* 2015;112:9442-7.
- 376 21. Nigro SJ, Hall RM. *GI_{sul2}*, a genomic island carrying the *sul2* sulphonamide
377 resistance gene and the small mobile element CR2 found in the *Enterobacter cloacae*
378 subspecies *cloacae* type strain ATCC 13047 from 1890, *Shigella flexneri* ATCC 700930 from
379 1954 and *Acinetobacter baumannii* ATCC 17978 from 1951. *J Antimicrob Chemother*
380 2011;66:2175-6.
- 381 22. Hamidian M, Venepally P, Hall RM, Adams MD. Corrected genome sequence of
382 *Acinetobacter baumannii* strain AB0057, an antibiotic-resistant isolate from lineage 1 of
383 global clone 1. *Genome Announc* 2017;5:e00836-17.
- 384 23. Hamidian M, Hawkey J, Wick R, Holt KE, Hall RM. Evolution of a clade of
385 *Acinetobacter baumannii* global clone 1, lineage 1 via acquisition of carbapenem- and

- 386 aminoglycoside-resistance genes and dispersion of ISAbal. *Microb Genom.*
387 2019;5:mgen.0.000242.
- 388 24. Hua X, Xu Q, Zhou Z, Ji S, Yu Y. Relocation of Tn2009 and characterization of an
389 ABGRI3-2 from resequenced genome sequence of *Acinetobacter baumannii* MDR-ZJ06. *J*
390 *Antimicrob Chemother* 2019;74:1153-1155.
- 391 25. Blackwell GA, Holt KE, Bentley SD, Hsu LY, Hall RM. Variants of AbGRI3
392 carrying the *armA* gene in extensively antibiotic-resistant *Acinetobacter baumannii* from
393 Singapore. *J Antimicrob Chemother* 2017;72:1031-9.
- 394 26. Adams MD, Goglin K, Molyneaux N, Hujer KM, Lavender H, Jamison JJ, *et al.*
395 Comparative genome sequence analysis of multidrug-resistant *Acinetobacter baumannii*. *J*
396 *Bacteriol* 2008;190:8053-64.
- 397 27. Iacono M, Villa L, Fortini D, Bordoni R, Imperi F, Bonnal RJ, *et al.* Whole-genome
398 pyrosequencing of an epidemic multidrug-resistant *Acinetobacter baumannii* strain belonging
399 to the European clone II group. *Antimicrob Agents Chemother* 2008;52:2616-25.
- 400 28. Russo TA, Luke NR, Beanan JM, Olson R, Sauberan SL, MacDonald U, *et al.* The
401 K1 capsular polysaccharide of *Acinetobacter baumannii* strain 307-0294 is a major virulence
402 factor. *Infect Immun* 2010;78:3993-4000.
- 403 29. Russo TA, Manohar A, Beanan JM, Olson R, MacDonald U, Graham J, *et al.* The
404 Response Regulator BfmR is a Potential Drug Target for *Acinetobacter baumannii*. *mSphere*
405 2016;1:e00082-16.
- 406 30. Vallejo JA, Beceiro A, Rumbo-Feal S, Rodriguez-Palero MJ, Russo TA, Bou G.
407 Optimisation of the *Caenorhabditis elegans* model for studying the pathogenesis of
408 opportunistic *Acinetobacter baumannii*. *Int J Antimicrob Agents* 2015:S0924-8579.

- 409 31. Wang-Lin SX, Olson R, Beanan JM, MacDonald U, Balthasar JP, Russo TA. The
410 Capsular polysaccharide of *Acinetobacter baumannii* Is an obstacle for therapeutic passive
411 immunization strategies. *Infect Immun* 2017;85:e00591-17.
- 412 32. Hamidian M, Ambrose SJ, Hall RM. A large conjugative *Acinetobacter baumannii*
413 plasmid carrying the *sul2* sulphonamide and *strAB* streptomycin resistance genes. *Plasmid*
414 2016;87-88:43-50.
- 415 33. Kenyon JJ, Hall RM. Variation in the complex carbohydrate biosynthesis loci of
416 *Acinetobacter baumannii* genomes. *PLoS One* 2013;8:e62160.
- 417 34. Kenyon JJ, Hall RM. Variation in the complex carbohydrate biosynthesis loci of
418 *Acinetobacter baumannii* genomes. *PLoS One* 2013;8:e62160.
- 419 35. Hamidian M, Hall RM. pACICU2 is a conjugative plasmid of *Acinetobacter* carrying
420 the aminoglycoside resistance transposon *TnaphA6*. *J Antimicrob Chemother* 2014;69:1146-
421 8.
- 422 36. Wilson K. Preparation of genomic DNA from bacteria. *Current protocols in Mol Biol*
423 2001;Chapter 2:Unit 2.4.
- 424 37. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*
425 2014;30:2068-9.
- 426 38. Hamidian M, Hall RM. The AbaR antibiotic resistance islands found in *Acinetobacter*
427 *baumannii* global clone 1 – structure, origin and evolution. *Drug Resist Updat* 2018;41:26-
428 39.
- 429 39. Nigro SJ, Hall RM. Tn6167, an antibiotic resistance island in an Australian
430 carbapenem-resistant *Acinetobacter baumannii* GC2, ST92 isolate. *J Antimicrob Chemother*
431 2012;67:1342-6.

- 432 40. Bertini A, Poirel L, Mugnier PD, Villa L, Nordmann P, Carattoli A. Characterization
433 and PCR-based replicon typing of resistance plasmids in *Acinetobacter baumannii*.
434 *Antimicrob Agents Chemother* 2010;54:4168-77.
- 435 41. Hamidian M, Holt KE, Pickard D, Dougan G, Hall RM. A GC1 *Acinetobacter*
436 *baumannii* isolate carrying AbaR3 and the aminoglycoside resistance transposon TnaphA6 in
437 a conjugative plasmid. *J Antimicrob Chemother* 2014;69:955-8.
- 438 42. Smith MG, Gianoulis TA, Pukatzki S, Mekalanos JJ, Ornston LN, Gerstein M, *et al.*
439 New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density
440 pyrosequencing and transposon mutagenesis. *Genes Dev* 2007;21:601-14.
- 441 43. Vallenet D, Nordmann P, Barbe V, Poirel L, Mangenot S, Bataille E, *et al.*
442 Comparative analysis of *Acinetobacters*: three genomes for three lifestyles. *PLoS One*
443 2008;3:e1805.
- 444 44. Park JY, Kim S, Kim SM, Cha SH, Lim SK, Kim J. Complete genome sequence of
445 multidrug-resistant *Acinetobacter baumannii* strain 1656-2, which forms sturdy biofilm. *J*
446 *Bacteriol* 2011;193:6393-4.
- 447 45. Chen CC, Lin YC, Sheng WH, Chen YC, Chang SC, Hsia KC, *et al.* Genome
448 sequence of a dominant, multidrug-resistant *Acinetobacter baumannii* strain, TCDC-AB0715.
449 *J Bacteriol* 2011;193:2361-2.
- 450 46. Zhou H, Zhang T, Yu D, Pi B, Yang Q, Zhou J, *et al.* Genomic analysis of the
451 multidrug-resistant *Acinetobacter baumannii* strain MDR-ZJ06 widely spread in China.
452 *Antimicrob Agents Chemother* 2011;55:4506-12.
- 453 47. Liou ML, Liu CC, Lu CW, Hsieh MF, Chang KC, Kuo HY, *et al.* Genome sequence
454 of *Acinetobacter baumannii* TYTH-1. *J Bacteriol* 2012;194:6974.
- 455 48. Gao F, Wang Y, Liu YJ, Wu XM, Lv X, Gan YR, *et al.* Genome sequence of
456 *Acinetobacter baumannii* MDR-TJ. *J Bacteriol* 2011;193:2365-6.

457

458 **11. Data bibliography**

- 459 1. **Adams, M.D., Goglin, K., Molyneaux, N., Hujer, K.M., Lavender, H., Jamison, J.J.,**
460 **MacDonald, I.J., Martin, K.M., Russo, T., Campagnari, A.A., Hujer, A.M., Bonomo, R.A.**
461 **and Gill, S.R.** NCBI GenBank CP012952 *A. baumannii* AB307-0294 (2008).
- 462 2. **Carattoli, A., Villa, L., Fortini, D. and Cassone, A.** NCBI GenBank CP000863.1 *A.*
463 *baumannii* ACICU, complete genome (2007).
- 464 3. **Hamidian, M., Wick, R., Judd, L., Russo, T.A., Holt, K.E. and Hall, R.M.** NCBI
465 GenBank CP012952 *A. baumannii* AB307-0294 (2017).
- 466 4. **Hartstein RM, Hamidian M, Nigro SJ, Wick R, Judd L, Holt K, Hall RM.** NCBI
467 GenBank CP031380.1 *A. baumannii* isolate ACICU (2019).
- 468 5. **Hua, X.** NCBI GenBank CP001937.2 *Acinetobacter baumannii* MDR-ZJ06, complete genome
469 (2018).
- 470 6. **Smith, M.G., Gianoulis, T.A., Pukatzki, S., Mekalanos, J.J., Ornston, L.N., Gerstein, M. and**
471 **Snyder, M.** NCBI GenBank CP000521.1 *A. baumannii* ATCC 17978, complete genome (2008).
- 472 7. **Weber, B.S., Ly, P.M., Irwin, J.N., Pukatzki, S. and Feldman, M.F.** NCBI GenBank
473 CP012004.1 *A. baumannii* ATCC 17978-mff, complete genome (2013).

474 **12. Figures and tables**

475 **12.1 Figure legends**

476 **Figure 1. Histograms of CDS lengths relative to the length of the top hit in Uniprot, in**
477 **the original vs revised genomes.** A) ACICU GenBank accession no. CP000863.1 (original)
478 and CP031380 (revised), B) AB307-0294 GenBank accession no. CP001172.1 (original) and
479 CP001172.2 (revised), and C) ATCC 17978 GenBank accession no. CP000521.1 (original)
480 and CP012004.1 (revised). The x-axis shows the ratio of coding sequence length to the length
481 of the closest hit in the UniProt TrEMBL database. The y-axis shows gene frequency and is
482 truncated at 100 (the centre bar extends to ~3000 genes). A tight distribution around 1.0
483 indicates that the assembly's coding sequences match known proteins, supporting few indel
484 errors in the assembly. A left-skewed distribution is characteristic of an assembly with indel
485 errors which lead to premature stop codons.

486

487

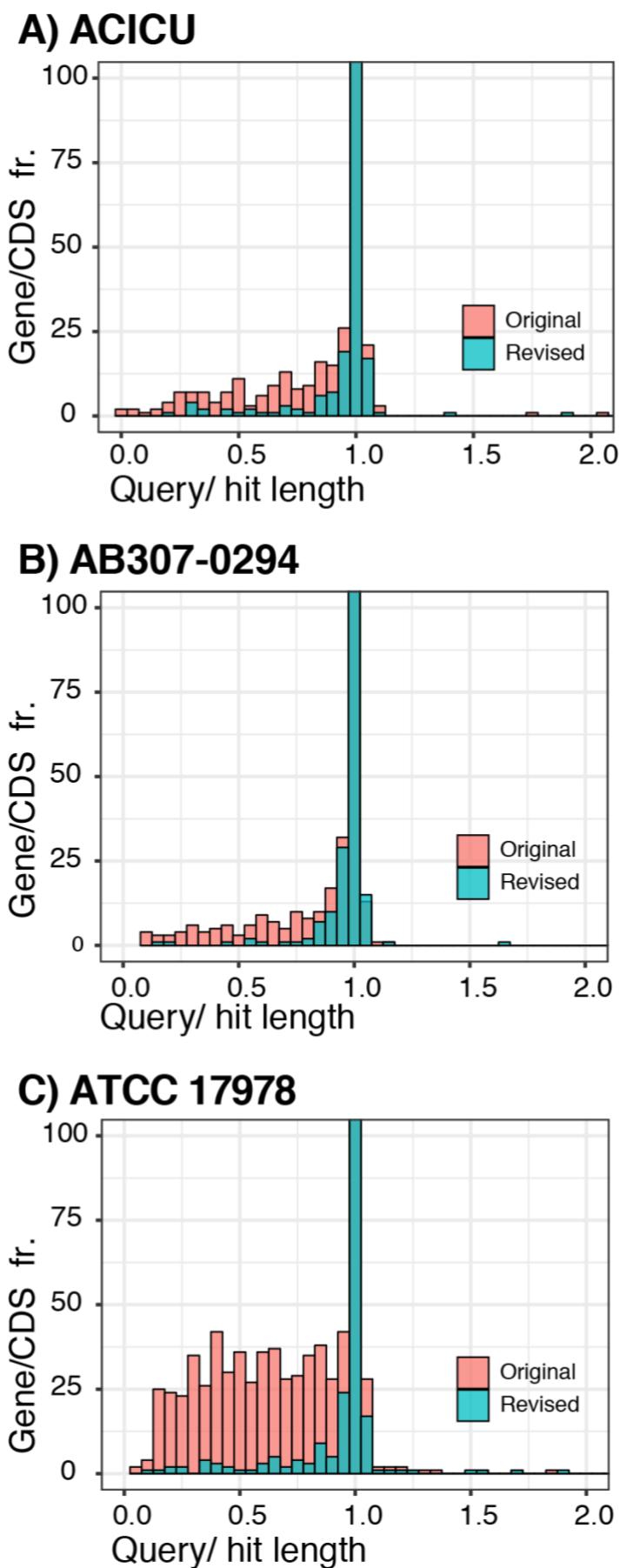
| | | | | | | | | | | | |
|-----------------------------|-------|-------------------|---|---------|------------|----------|------|----|---|---|---|
| Chromosome MDR-TJ | China | 2012 ^c | 2 | 3957368 | CP003856 | Illumina | No | - | - | - | - |
| Chromosome pABTJ1 | | | | 3964912 | CP003500.1 | 454 | (48) | No | - | - | - |
| pABTJ1 | | | | 77528 | CP003501.1 | " | | No | - | - | - |
| pABTJ1 | | | | 110967 | CP004359.1 | " | | No | - | - | - |

489 ^a nk: not known, na: not applicable.

490 ^b Global Clones.

491 ^c Genome submission date; isolation date is not known.

492 ^d recovered between 2007-2009



1 Figure 1.