

1 **The GMI-K70 collection: precisely annotated genomic islands, integration sites,**
2 **and a novel tRNA gene nomenclature across the *Klebsiella* genus**

3
4 Camilo Berríos-Pastén^{1,*}, Rodolfo Acevedo¹, Carlos Serrano-Pinto¹, Patricio Arros¹, Margaret M.
5 C. Lam^{2,3}, Kathryn E. Holt^{2,4}, Rosalba Lagos¹, Kelly L. Wyres^{2,3,4}, Andrés E. Marcoleta^{1,*}

6
7 ¹Grupo de Microbiología Integrativa, Laboratorio de Biología Estructural y Molecular BEM, Departamento de Biología,
8 Facultad de Ciencias, Universidad de Chile. Santiago 7800003, Chile.

9 ²Department of Infectious Diseases, School of Translational Medicine, Monash University, Melbourne 3004, Australia.

10 ³ Centre to Impact AMR, Monash University, Clayton 3800, Australia

11 ⁴ Department of Infection Biology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK.

12
13 Camilo Berríos-Pastén, e-mail: camilo.berrios.p@ug.uchile.cl

14 Andrés E. Marcoleta, e-mail: amarcoleta@uchile.cl

15
16 **Keywords:** *Klebsiella*, mobile genetic elements, chromosomal mobilome, prophages, genomic
17 islands, comparative genomics.

18 **Repositories:** The GenBank format files corresponding to the annotated and curated genomes
19 of the *K. pneumoniae* strains forming part of the GMI-K70 collection are accessible in the following
20 Zenodo entry: 10.5281/zenodo.20767732.

21 **ABSTRACT**

22 *Klebsiella pneumoniae* is a critical Gram-negative pathogen that frequently acquires and
23 traffics mobile genetic elements (MGEs) carrying virulence and antimicrobial resistance genes.
24 While plasmids have been extensively studied, chromosomal MGEs, particularly genomic islands
25 (GIs), and their contribution to *K. pneumoniae* evolution remain comparatively underexplored. GIs
26 commonly integrate into tRNA and tmRNA genes (t(m)DNAs) and play key roles in pathogenesis
27 across diverse bacterial phyla. However, progress in understanding these elements has been
28 hindered by the absence of a systematic nomenclature for identifying and comparing equivalent
29 t(m)DNA loci and by limited insight into the determinants governing integration site selection, even
30 among identical tDNA copies within the chromosome. To address these limitations, we
31 established the GMI-K70 collection, a curated dataset of 70 high-quality complete *Klebsiella* spp.
32 genomes, annotated with their complete repertoire of chromosomal GIs and integration sites.
33 Using a novel t(m)DNA nomenclature based on conserved genomic context, we achieved

34 unambiguous differentiation of identical t(m)DNAs across chromosomes, enabling precise
35 comparative synteny analyses for GI identification. In total, we identified 676 chromosomal MGEs
36 (median 10 per genome), substantially exceeding the GI burden previously detected by
37 automated approaches. Of these, 427 GIs were associated with 21 t(m)DNA loci, while 249 were
38 integrated at 46 non-t(m)DNA sites located downstream of coding sequences or within intergenic
39 regions. Notably, the *icd* gene emerged as a previously unrecognized integration hotspot, with a
40 usage frequency comparable to reported t(m)DNA sites such as *asn1C* and *phe1A*. Genome-
41 wide integrase gene detection mapped nearly all these genes within the predicted MGEs,
42 supporting the completeness of GI identification. The GMI-K70 collection provides a gold-
43 standard framework for benchmarking GI and prophage detection tools and offers a refined view
44 of chromosomal MGE diversity and integration dynamics in *Klebsiella*. This resource establishes
45 a foundation for future studies of genome evolution and the dissemination of virulence and
46 antimicrobial resistance determinants in this critical priority pathogen.

47 **IMPACT STATEMENT**

48 Mobile genetic elements are key drivers of genome evolution, virulence, and antimicrobial
49 resistance in *K. pneumoniae*. However, chromosomal mobile elements including genomic islands
50 (GIs), remain comparatively underexplored due to the absence of a systematic framework for
51 identifying integration sites across genomes. Here, we introduce the GMI-K70 collection, a
52 manually curated reference dataset that enables precise mapping of genomic islands and their
53 chromosomal integration sites across the *K. pneumoniae* species complex. By establishing a
54 standardized t(m)DNA nomenclature and providing a comprehensive catalogue of integration loci,
55 this resource supports accurate benchmarking of genomic island and prophage detection tools
56 and advances our understanding of chromosomal genome plasticity in this clinically important
57 pathogen.

58 **DATA SUMMARY**

59 The authors confirm that all supporting data, code, and protocols have been provided
60 within the article or through supplementary data files.

61 INTRODUCTION

62 *Klebsiella pneumoniae* is ranked as the top priority pathogen on the World Health
63 Organization's list of critical Gram-negative threats, driven by the global rise of multidrug-resistant
64 and hypervirulent lineages (1,2). This species acts as both a major reservoir and an efficient
65 disseminator of antibiotic resistance genes (ARGs), facilitating their spread to other clinically
66 important bacteria (3–5). Genomic studies have shown that virulence and resistance traits in
67 hypervirulent and multidrug-resistant strains are largely encoded by horizontally acquired genes
68 carried on diverse mobile genetic elements (MGEs). Beyond plasmids, genomic islands (GIs) in
69 *K. pneumoniae* harbour extensive repertoires of virulence and antimicrobial resistance
70 determinants (6–8).

71 GIs are discrete DNA segments variably present at equivalent chromosomal positions
72 among strains of the same or related species. Under certain conditions, some excise to form
73 circular intermediates capable of integrating into new host genomes, typically at tRNA genes
74 (tDNAs) or the transfer-messenger RNA gene (*ssrA*) (9–11), here collectively referred to as
75 “t(m)DNAs”. These elements can carry tens to hundreds of cargo genes. Integrative conjugative
76 elements (ICEs), for example, encode a type IV secretion system and an origin of transfer (*oriT*)
77 enabling conjugative dissemination, whereas mobilizable GIs encode only an *oriT* and rely on
78 related conjugative elements for transfer (12).

79 Most GIs and prophages encode an integrase that mediates site-specific recombination
80 between the chromosomal *attB* site, usually corresponding to a short sequence located at the 3'
81 end of the target t(m)DNA, and the corresponding *attP* site within the element, generating flanking
82 direct repeats upon integration (9). Although integrase specificity is considered a primary
83 determinant of integration site selection, the reasons why some t(m)DNAs are preferentially
84 targeted remain unresolved, and the extent to which alternative chromosomal loci function as
85 integration sites is unclear (13). Moreover, to our knowledge, no studies have systematically
86 examined the complete set of t(m)DNAs across multiple *K. pneumoniae* chromosomes to
87 evaluate their properties as integration sites.

88 Previous work in *K. pneumoniae* identified 12 GI families integrating into asparagine
89 tDNAs, establishing these loci as major integration hotspots (6). These islands include GIE492,
90 which encodes the determinants for producing the antibacterial peptide microcin E492, and
91 multiple ICE*Kp* variants that carry genes for production of the yersiniabactin siderophores and the
92 colibactin genotoxin (7). Co-occurrence of GIE492 and ICE*Kp10* is strongly associated with

93 hypervirulent *K. pneumoniae* (hvKp), and both microcin E492 and colibactin synergistically
94 enhance gut colonization and persistence (14), underscoring the central role of GIs in
95 pathogenicity.

96 Notably, although *K. pneumoniae* chromosomes typically carry four identical copies of the
97 asparagine tDNA, their usage as integration sites differs markedly (6,7). This observation
98 suggests that additional genomic or molecular factors influence site selection. Questions remain
99 to be answered, such as whether similar biases occur in other t(m)DNAs, whether all t(m)DNAs
100 can serve as integration sites, and whether unidentified chromosomal loci contribute to the
101 diversity of GI. It is therefore necessary to carry out a systematic analysis of the entire genome to
102 identify integration sites in *Klebsiella* species, and this is expected to reveal a considerable
103 number of previously unknown GIs.

104 A major obstacle to such analyses is the absence of a standardized t(m)DNA
105 nomenclature. Current annotation tools report only anticodon identity and encoded amino acid,
106 making it difficult to distinguish identical tDNA copies located in distinct genomic contexts. A
107 contextual nomenclature integrating both anticodon identity and conserved flanking gene
108 architecture was recently proposed for asparagine tDNAs (6), but has not been extended
109 genome-wide.

110 To address these limitations, we established the GMI-K70 collection, a curated dataset of
111 70 high-quality complete *Klebsiella* sp. chromosome sequences with comprehensive annotations
112 of integrative mobile genetic elements and loci used as integration sites. We extended the
113 t(m)DNA gene nomenclature based on conserved genomic context, enabling genome-wide and
114 unambiguous identification of identical tDNAs. Applying this framework, we systematically curated
115 676 chromosomal MGEs and identified numerous previously unrecognized integration sites,
116 including loci located downstream of protein-coding genes and within intergenic regions.

117 By providing a comprehensive map of integration sites, precisely delineated GI
118 boundaries, and integrase annotations, the GMI-K70 collection serves as a gold-standard
119 reference for benchmarking MGE detection tools. More broadly, this resource establishes a
120 foundation for investigating chromosome evolution and the dissemination of virulence and
121 antimicrobial resistance determinants across clinically important *Klebsiella* lineages.

122 **METHODS**

123 **Source of DNA sequences and phylogenomic analysis**

124 At the outset of this strain annotation project in March 2015, nearly 100 complete *K.*
125 *pneumoniae* assemblies were available in the NCBI database. Genome completeness and
126 contamination were assessed using CheckM v2 (15), and only those meeting completeness $\geq 95\%$
127 and contamination $\leq 5\%$ criteria were retained for further analysis. Plasmid sequences were
128 identified and removed, and all subsequent analyses were performed exclusively on
129 chromosomal sequences. Accession numbers and associated metadata are listed in Table S1.

130 To ensure accurate taxonomic classification, species assignments and sequence types
131 (STs) were determined using Kleborate v3.1.0 (16). Sublineages and clonal groups based on the
132 LIN codes and the scgMLSTv2 scheme (17) were subsequently assigned using the tools available
133 on the PathogenWatch platform (18). Following taxonomic curation, the final dataset comprised
134 70 high-quality complete genomes belonging to *Klebsiella* spp., including 39 *K. pneumoniae*, 17
135 *K. quasipneumoniae*, 8 *K. variicola*, 3 *K. michiganensis*, 2 *K. oxytoca*, and 1 *K. aerogenes*. This
136 curated strain collection was designated GMI-K70.

137 Phylogenetic relationships among strains were inferred by first performing a pangenome
138 analysis with Panaroo v1.5.0 (19) using default parameters (core genes defined as those present
139 in $\geq 95\%$ of strains). A core genome multiple sequence alignment was subsequently generated
140 and used to infer a maximum-likelihood phylogeny in IQ-TREE v2.3.0 (20) under the GTR+G
141 substitution model, with branch support estimated using 1,000 ultrafast bootstrap replicates.

142 **Identification and nomenclature of t(m)DNAs**

143 Initial identification of t(m)DNAs was performed using ARAGORN v1.2.38 (21) and
144 tRNAscan-SE v2.0.12 (22), which predict anticodon sequence, genomic coordinates, and
145 transcriptional orientation. To analyse conserved genomic contexts, annotations generated with
146 Bakta v1.9.2 (23) (database v5.1) were manually curated using the SnapGene software v7.2.0
147 (www.snapgene.com). This strategy enabled the identification of conserved coding sequences
148 (CDSs) located upstream and downstream of each t(m)DNA locus.

149 The proposed nomenclature system is based on conserved genomic context, specifically
150 the three CDSs immediately upstream and three immediately downstream of each t(m)DNA,
151 defined operationally. The presence of neighbouring t(m)DNA and rRNA genes was also
152 considered. Thus, the nomenclature integrates three elements: the encoded amino acid, the

153 anticodon sequence (as defined in Table S2), and a contextual identifier derived from the
154 conserved flanking region.

155 **Identification of genomic islands**

156 In this study, genomic islands (GIs) were operationally defined as DNA segments
157 exceeding 2.5 kbp that disrupt the synteny of otherwise conserved chromosomal regions across
158 strains. The term “genomic islands” is used here broadly to encompass integrative mobile
159 elements, including prophages, integrative and conjugative elements (ICEs), and classical
160 genomic islands (24). Although functionally distinct, these elements were analysed collectively as
161 a unified class of chromosomal integrative elements, as growing evidence supports a shared
162 evolutionary origin and overlapping mechanistic features among these integrative mobile
163 elements (9,25,26). To identify t(m)DNA-associated MGEs, conserved genomic anchors flanking
164 each t(m)DNA were used as reference points. Insertions disrupting these conserved contexts
165 were manually delineated. Direct repeats (DRs) were detected using BLASTn v2.16.0 (27), and
166 integrase-encoding genes were identified through BLASTp v2.16.0 searches against a custom
167 database of bacterial and phage integrase proteins. This database was constructed using
168 sequences retrieved from UniProt (accessed in September 2024), with identification thresholds
169 set at a minimum of 90% identity and 80% coverage.

170 To expand detection beyond t(m)DNA loci, conserved syntenic blocks defined by single-
171 copy core genes were used to identify accessory regions interrupting otherwise conserved
172 chromosomal architecture. Candidate regions were further evaluated for the presence of DRs,
173 integrase genes, and deviations in GC content relative to the host chromosome. This strategy
174 enabled systematic identification of MGEs integrated within coding sequences (CDSs) and
175 intergenic regions.

176 **Comparative analysis of t(m)DNA sequences**

177 t(m)DNA loci were extracted from chromosomal assemblies and aligned using ClustalW
178 v2.1 (28). Sequences were grouped according to the proposed nomenclature, and pairwise
179 nucleotide identity values were calculated from the alignments. Prevalence was defined as the
180 proportion of strains containing at least one copy of a given t(m)DNA type.

181 **Clustering of genomic islands based on nucleotide sequence and functional annotation of** 182 **MGEs**

183 MGE nucleotide sequences were extracted from chromosomal assemblies. For elements
184 with identifiable DRs, these were used to define boundaries. In the absence of DRs, boundaries

185 were defined as extending to the 3' end of the intergenic region between the integration site and
186 the first flanking conserved gene.

187 MGEs were clustered using MMSeqs2 v16.747c6 (29) with thresholds of $\geq 85\%$ nucleotide
188 identity and $\geq 80\%$ coverage to account for structural variation arising from insertion sequences
189 and transposons. Clustering was performed both within individual integration sites and across the
190 entire MGE dataset to identify elements shared between loci.

191 Functional annotation was based on Bakta-generated gene predictions. Protein-coding
192 sequences were compared against the Virulence Factor Database (VFDB 2022) (30) to identify
193 virulence-associated genes, analysed with AMRFinderPlus v4.0.3 (31) to detect antimicrobial
194 resistance genes, and assigned to COG functional categories using COGclassifier v1.0.5
195 (<https://github.com/moshi4/COGclassifier>). Prophage regions were predicted using PHASTEST
196 (32), and coordinates were compared with curated GI boundaries.

197

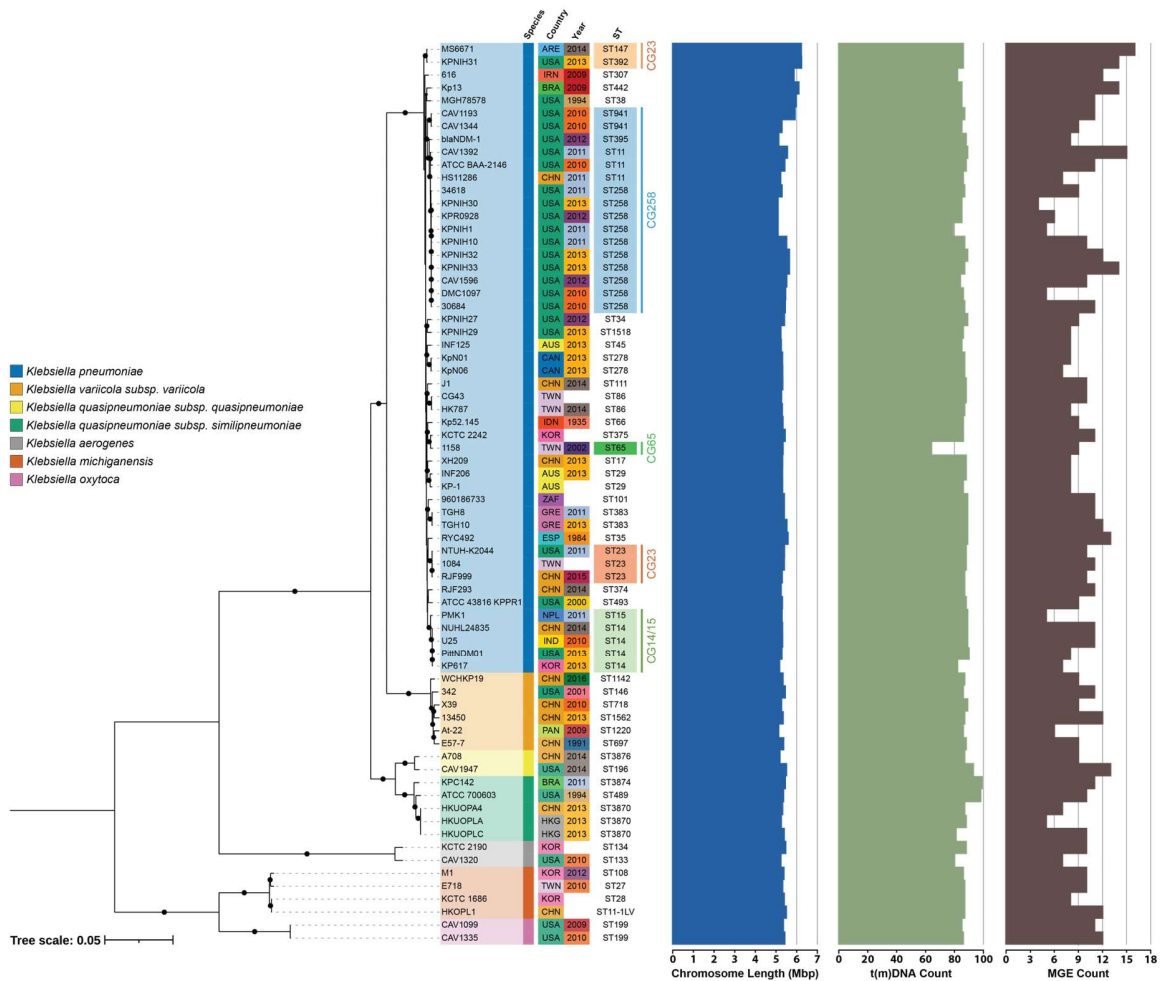
198 **RESULTS AND DISCUSSION**

199

200 **The GMI-K70 collection captures the phylogenomic breadth of the *Klebsiella* genus**

201 To enable a comprehensive search for putative genomic islands (GIs) and related mobile
202 elements targeting t(n)DNAs, we first established a phylogenetically diverse and taxonomically
203 curated genome dataset. The GMI-K70 collection encompasses representative members of the
204 *Klebsiella* genus, including *K. pneumoniae*, *K. variicola*, *K. quasipneumoniae*, *K. aerogenes*, *K.*
205 *michiganensis*, and *K. oxytoca*. In total, the dataset represents 45 distinct sequence types, spans
206 a broad temporal range of isolation (1935–2016), and includes strains from multiple geographic
207 regions.

208 Phylogenomic analysis revealed clear species-level clustering and extensive diversity
209 within *K. pneumoniae* sensu stricto, encompassing major global clonal groups such as CG258,
210 CG147, CG15, CG23, and CG485 (Fig. 1). Other members of the genus such as *K. michiganensis*
211 and *K. oxytoca* formed well-supported, deeply branching lineages, underscoring the broad
212 phylogenetic representation captured in the dataset. Collectively, the phylogeny spans the
213 *Klebsiella* evolutionary breadth, ensuring robust representation of both deep species divergence
214 and clinically relevant clonal structure. This phylogenetic framework provides a solid foundation
215 for evaluating integration site conservation and diversity across the chromosome of species from
216 this genus.



217
218

219 **Figure 1. Phylogenetic diversity and genomic characteristics of the GMI-K70 collection.** Maximum-likelihood
220 phylogeny of the 70 genomes included in the GMI-K70 collection, inferred from a core genome alignment of 3,186
221 genes present in $\geq 95\%$ of strains. The tree illustrates broad phylogenetic representation across *Klebsiella*, including
222 major globally distributed clinical clonal groups. Species assignments and sequence types (STs) were determined using
223 Kleborate. Black circles indicate nodes with ultrafast bootstrap support $\geq 80\%$. Annotated panels display metadata for
224 year and country of isolation, as well as genomic features for each strain, including chromosome length (Mbp), total
225 number of t(m)DNA loci, and number of curated chromosomal MGEs.

226

227 **A conserved-context nomenclature defines core and accessory t(m)DNAs in *Klebsiella***

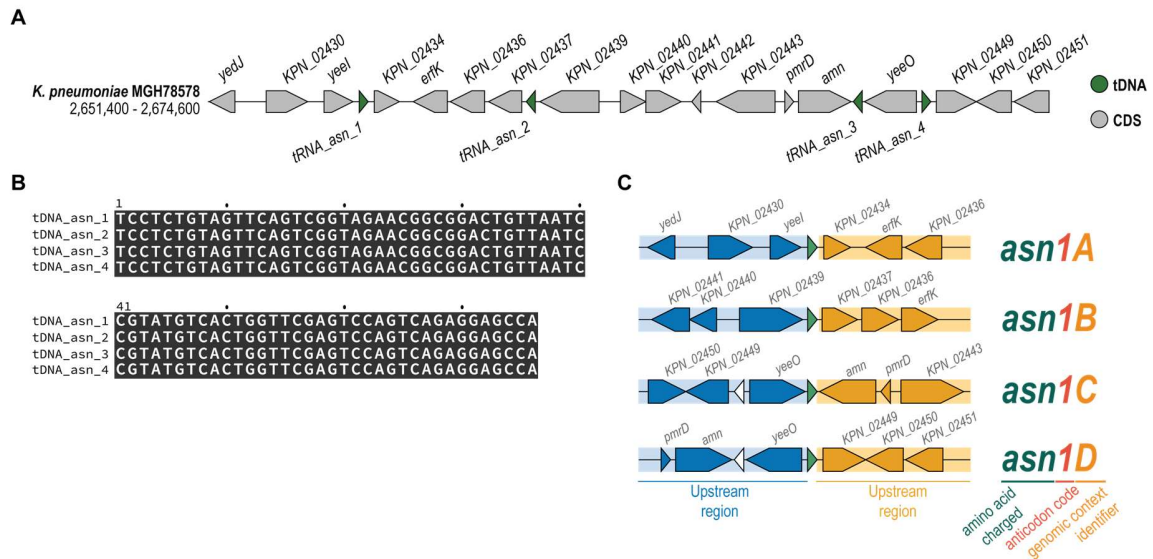
228 Because several tDNAs occur in multiple, often identical copies within *Klebsiella*
229 chromosomes, we developed a novel nomenclature capable of unambiguously resolving these
230 copies. For example, the four asparagine tDNAs (*asn1A*, *asn1B*, *asn1C*, and *asn1D*) share 100%
231 nucleotide identity yet occupy distinct chromosomal positions with different flanking gene contexts

232 (Fig. 2). Existing positional nomenclature systems and automated annotation tools cannot reliably
233 distinguish these copies. By incorporating conserved genomic context as a defining criterion, our
234 nomenclature resolves this ambiguity and enables robust cross-genome comparisons. This
235 system was applied consistently to all chromosomal t(m)DNAs in the GMI-K70 collection (Fig.
236 S1). To define these genomic contexts, we evaluated three CDSs upstream and three CDSs
237 downstream of each t(m)DNA. This window size was selected empirically, as it represented the
238 minimum number of flanking genes required to successfully differentiate identical copies across
239 genomes during the manual curation of chromosomal annotations. An automated pipeline to
240 identify distinct t(m)DNA copies based on this nomenclature is currently under development.
241 Presently, the genomic context sequences provided in the accompanying repository can be
242 utilized for comparisons using alignment tools such as BLAST. For instance, a previous study by
243 another research group described that the genomic island (GI) ICEKp258.2 integrates specifically
244 into the *asn1D* copy of asparagine tDNAs, utilizing the surrounding flanking genes for
245 classification purposes (33).

246 Supporting this framework, we confirmed that, in the absence of integrated elements, the
247 ~4 kbp regions immediately upstream and downstream of each t(m)DNA are highly conserved
248 across *Klebsiella* chromosomes. These flanking regions, typically comprising three coding
249 sequences on each side, were sufficient to uniquely distinguish all t(m)DNA loci within the
250 chromosome. Once defined, these conserved genomic anchors facilitate comparative analyses
251 by enabling straightforward detection of integrated MGEs that disrupt this architecture (Fig. S2).

252 Across the 70 strains analysed, we identified 6,097 t(m)DNAs. All loci were manually
253 curated to determine their conserved genomic context and were classified into 157 distinct types
254 following the proposed nomenclature. The tmDNA (*ssrA*) occurred in a single conserved context,
255 and all strains encoded exactly one copy. A prevalent core set of 87 t(m)DNA types (86 tDNAs
256 and one tmDNA) was present in $\geq 75\%$ of strains. Each core type occupied a characteristic and
257 highly conserved genomic context and, collectively, the core set was distributed across 42
258 chromosomal regions (Fig. S1).

259 The remaining 73 tDNA types occurred in fewer than 18% of strains and were therefore
260 classified as rare. These rare loci were typically located in regions suggestive of assembly
261 artefacts or embedded within mobile genetic elements, as further discussed below. Nucleotide
262 identity within each t(m)DNA type of the core set was consistently high (98–100%), further
263 supporting the robustness of the nomenclature.



264

265 **Figure 2. Conserved-context t(m)DNA nomenclature enables discrimination of identical loci in distinct genomic**

266 **positions.** (A) The four asparagine tDNAs present in the chromosome of *K. pneumoniae* MGH78578 are shown as an

267 illustrative example. (B) Despite occupying different chromosomal positions, these four copies share 100% nucleotide

268 identity, as demonstrated by DNA sequence alignment. (C) To enable unambiguous differentiation of identical t(m)DNA

269 copies within and across genomes, their conserved genomic context was incorporated into the nomenclature. The

270 proposed naming system integrates three components: the transported amino acid, the encoded anticodon, and a

271 contextual identifier derived from the conserved flanking gene architecture.

272

273 Leveraging conserved t(m)DNA genomic context for chromosome-wide identification of 274 integrated MGEs

275 Using the conserved genomic architecture flanking each t(m)DNA as reference anchors,

276 we manually inspected these regions across the GMI-K70 chromosomes to identify insertion

277 events disrupting conserved synteny. This manual curation of t(m)DNA genomic contexts enabled

278 the identification of 427 chromosomal MGEs integrated within these loci.

279

280 In most cases, integration occurred at the 3' region of the t(m)DNA, consistent with the

281 canonical GI integration mechanism. Notably, the *arg2A* locus displayed an atypical pattern:

282 integration events were detected in the 5' region of the gene, and in some strains two distinct GIs

283 were integrated within the same *arg2A* locus, one upstream and one downstream. Whether this

284 configuration represents a unique feature of *arg2A* or reflects a broader but previously

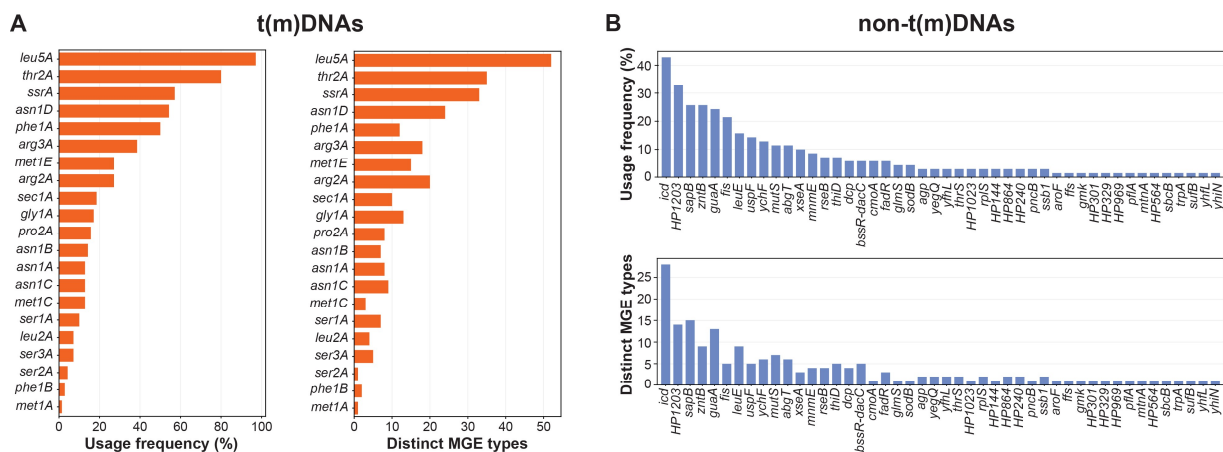
285 underappreciated integration pattern among t(m)DNAs will require analysis of larger genome

collections.

286 In total, 21 of the 87 core t(m)DNA types were associated with MGEs in at least one strain.
 287 Integration site usage was highly uneven across loci. The most frequently targeted sites
 288 corresponded to previously recognized hotspots: *leu5A* (100% integration frequency), *thr2A*
 289 (85.7%), *ssrA* (61.4%), *asn1D* (60.0%), and *phe1A* (52.9%) (Fig. 3A). These findings confirm that
 290 only a subset of core t(m)DNAs act as recurrent and preferential integration targets, while others
 291 remain rarely or never occupied.

292 Collectively, these results demonstrate that the conserved genomic architecture
 293 surrounding t(m)DNAs provides a robust framework for systematic, chromosome-wide
 294 identification of integrated MGEs. By exploiting conserved flanking regions as stable genomic
 295 anchors, this strategy enables precise detection of integration events and accurate delineation of
 296 MGE boundaries, including elements that may escape detection by composition-based or fully
 297 automated prediction methods.

298
 299



300
 301
 302 **Figure 3. Chromosomal integration site usage and diversity of mobile genetic elements in the GMI-K70**
 303 **collection.** (A) Frequency of MGE integration at tRNA- and tmRNA-encoding loci across the GMI-K70 genomes. (B)
 304 Number of distinct MGEs identified at each t(m)DNA locus. (C) Frequency of MGE integration at non-t(m)DNA
 305 chromosomal loci, including protein-coding genes and intergenic regions. (D) Number of distinct MGEs identified at
 306 each non-t(m)DNA integration site. Distinct MGEs were defined by clustering nucleotide sequences using MMSeqs2
 307 with a $\geq 85\%$ identity threshold and $\geq 80\%$ coverage. For integration events occurring within coding sequences (CDSs)
 308 annotated as hypothetical proteins, protein length was used as an auxiliary criterion for locus differentiation and naming
 309 (e.g., “HP1203” denotes a CDS encoding a 1,203-amino-acid protein).

310 **A variety of unrecognized non-t(m)DNA loci act as GI integration sites**

311 Although exhaustive curation of MGEs integrated at t(m)DNA loci captured a substantial
312 fraction of the chromosomal accessory genome, it remained unclear whether these sites
313 accounted for the full repertoire of integrated elements. To address this, we examined the
314 chromosomal pangenome and observed that a considerable proportion of accessory genes were
315 not associated with t(m)DNA-linked MGEs. This discrepancy suggested the occurrence of
316 additional integration sites elsewhere in the chromosome. We therefore hypothesized that non-
317 t(m)DNA loci could also act as recurrent GI integration targets and applied a strategy analogous
318 to that used for t(m)DNAs to systematically identify these sites, quantify their usage frequency,
319 and characterize the associated MGEs.

320 Guided by this hypothesis, we extended our analysis across the remaining chromosomal
321 regions to detect signatures of integrated MGEs. Conserved syntenic blocks defined by single-
322 copy core genes were used as genomic anchors to identify accessory regions disrupting
323 otherwise conserved chromosomal architecture, indicative of insertion events. This approach
324 enabled us to distinguish bona fide MGEs from regions shaped by homologous recombination or
325 assembly artefacts and to confidently map their integration sites.

326 Using this framework, we identified 46 distinct non-t(m)DNA loci that served as integration
327 sites for MGEs. These elements were predominantly inserted at the 3' ends of protein-coding
328 genes or within intergenic regions, consistent with site-specific integration mechanisms. Overall,
329 integration frequency at these loci was lower than that observed for t(m)DNAs, indicating that
330 t(m)DNAs remain the principal chromosomal targets for GI integration in *Klebsiella*.

331 A notable exception was the *icd* gene, encoding isocitrate dehydrogenase, which emerged
332 as the most frequently used non-t(m)DNA integration site. Its usage frequency was comparable
333 to that of well-established t(m)DNA hotspots such as *asn1C* and *phe1A*, identifying *icd* as a
334 previously unrecognized major integration hotspot. MGEs integrated at this locus exhibited
335 structural and functional diversity comparable to those associated with canonical t(m)DNA loci,
336 supporting its role as a recurrent and biologically relevant target (Fig. 3B). In total, 67 distinct
337 chromosomal integration sites were identified across the dataset, comprising 21 t(m)DNA loci and
338 46 protein-coding or intergenic loci.

339 Overall, these findings demonstrate that while t(m)DNAs account for a substantial
340 proportion of chromosomally integrated MGEs, a diverse set of non-t(m)DNA loci also contributes
341 meaningfully to accessory genome composition. Incorporating these additional integration sites

342 provides a more complete and accurate view of GI distribution and integration dynamics across
343 *Klebsiella* chromosomes.

344 **Structural and compositional features of chromosomal mobile genetic elements in the** 345 **GMI-K70 collection**

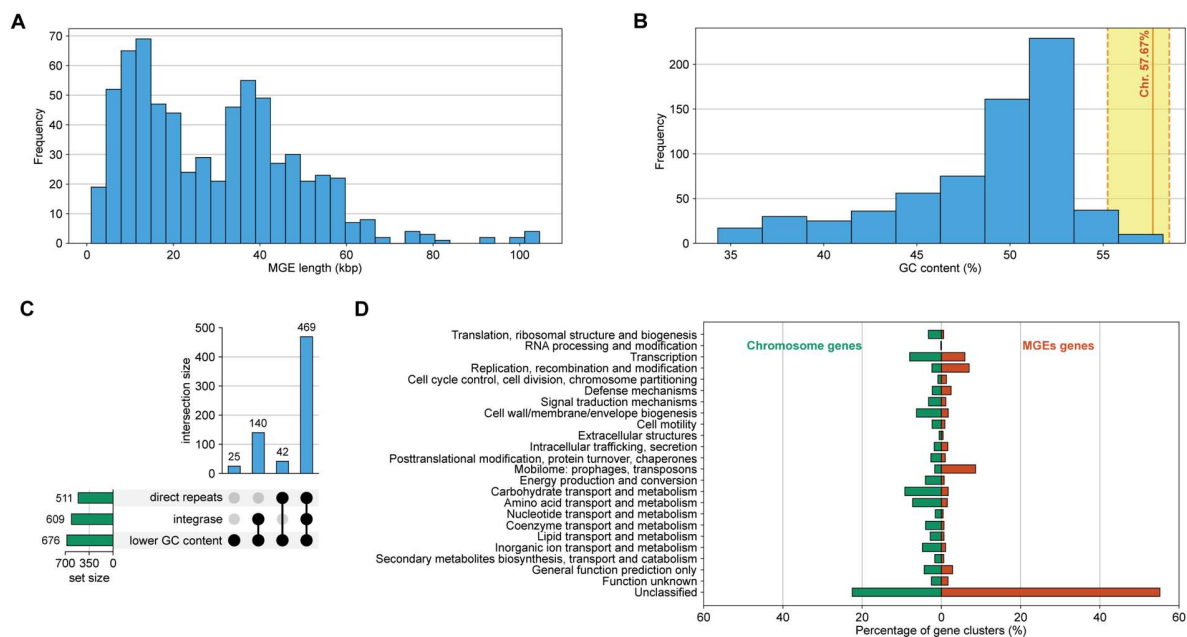
346 Across the 70 genomes of the GMI-K70 collection, we identified 676 chromosomal MGEs.
347 Of these, 427 were integrated at t(m)DNA loci, whereas 249 were observed at non-t(m)DNA sites.
348 This distribution corresponds to a median of 10 GIs per strain, slightly exceeding the median of 9
349 GIs per strain reported by the TIGER & Islander database (v1.0) for the subset of GMI-K70
350 isolates represented in that resource (27).

351 The identified MGEs exhibited substantial size variability, with an average length of 29,392
352 bp (range: 2,379–104,636 bp) (Fig. 4A). Their GC content averaged 48.66%, spanning 34.31%
353 to 58.22%. While many elements displayed pronounced compositional divergence from the host
354 chromosome, a subset exhibited GC contents closely matching the chromosomal background
355 (Fig. 4B). These cases illustrate a key limitation of GI detection methods that rely primarily on
356 nucleotide composition bias. Regarding integration signatures, 165 MGEs lacked detectable
357 direct repeats. This absence may reflect erosion or mutation of repeat sequences over time,
358 effectively stabilizing the elements within the chromosome, or may result from assembly
359 limitations, as repetitive regions remain challenging to resolve accurately.

360 All integrase-coding genes detected in the chromosomes were manually examined, and
361 the corresponding proteins were compared against integrase sequences present in UniProt. Most
362 integrases identified in the GMI-K70 genomes were encoded within curated MGEs; only 67 MGEs
363 lacked a detectable integrase gene. Integrases located outside annotated MGEs corresponded
364 primarily to the *fimB* and *fimE* genes, as well as a small number of solitary integrase genes
365 positioned downstream of the *arg2A*, *arg3A*, and *leu5A* loci. These latter cases likely represent
366 remnants of past integration events, consistent with recurrent insertion and subsequent erosion
367 at these chromosomal sites.

368 Functional annotation of the 676 MGEs revealed a gene repertoire sharply distinct from
369 that of the core chromosome (Fig. 4D). Whereas chromosomal genes were enriched in functions
370 associated with essential cellular processes (e.g., information processing, metabolism, and
371 cellular maintenance), MGE-encoded genes were strongly biased toward categories linked to
372 genome plasticity and ecological adaptation. In particular, MGEs were markedly enriched in genes
373 involved in replication, recombination, and repair, reflecting the high prevalence of integrases,

374 recombinases, and transposition-related proteins. Additional overrepresented categories included
 375 DNA defence mechanisms, signal transduction, and genes of unknown function, the latter
 376 highlighting the substantial fraction of uncharacterized cargo carried by these elements.
 377 Conversely, functions associated with core metabolic pathways were largely absent from MGEs.
 378



379
 380 **Figure 4. Structural and functional characteristics of chromosomal MGEs in the GMI-K70 collection.** (A) Size
 381 distribution of the 676 curated chromosomal MGEs. (B) GC content distribution of MGEs. For comparison, the average
 382 GC content of the corresponding host chromosomes is shown, with dashed lines indicating the minimum (55.2%) and
 383 maximum (58.56%) chromosomal GC content observed across the dataset. (C) UpSet plot summarizing the co-
 384 occurrence of key structural features among MGEs, including presence or absence of integrase genes and flanking
 385 direct repeats (DRs). (D) Proportional distribution of COG functional categories among gene clusters identified in host
 386 chromosomes and in MGEs. A total of 21,891 gene clusters were identified in chromosomal sequences, and 8,577
 387 gene clusters were identified within MGEs.

388
 389 Taken together, the near-universal association of integrases with curated MGEs, the
 390 systematic mapping of integration sites, and the distinct functional composition of MGE cargo
 391 provide strong support for the completeness and internal consistency of the GMI-K70 collection.
 392 These results reinforce the central role of chromosomal MGEs in shaping genome architecture
 393 and functional diversification in *Klebsiella*.

394 **CONCLUSIONS**

395 In this study, we established the GMI-K70 collection, a curated and phylogenetically
396 representative dataset that enables high-resolution characterization of chromosomal mobile
397 genetic elements across *Klebsiella*. By introducing a conserved-context t(m)DNA nomenclature,
398 we achieved unambiguous identification of identical tDNA copies and systematically defined their
399 associated integration sites. This framework revealed 676 chromosomal MGEs distributed across
400 67 distinct loci, including canonical t(m)DNA hotspots and 46 previously underappreciated non-
401 t(m)DNA sites, such as the highly targeted *icd* locus.

402 The conserved genomic anchors flanking t(m)DNAs, together with the high sequence
403 identity within defined t(m)DNA types, underscore the robustness and transferability of the
404 proposed nomenclature. Coupled with manual curation of integration boundaries and integrase
405 assignments, the GMI-K70 collection constitutes a high-quality reference for chromosomal MGE
406 architecture in *Klebsiella* genomes.

407 As a systematically curated resource, GMI-K70 enables benchmarking of GI and
408 prophage detection tools and provides a foundation for investigating chromosome evolution and
409 the horizontal dissemination of virulence and antimicrobial resistance determinants in *Klebsiella*.

410

411 **SUPPLEMENTARY MATERIAL**

412

413 **Figure S1. Conserved genomic architecture of core t(m)DNA loci in *Klebsiella* spp.**
414 **chromosomes.** The 42 chromosomal regions encompassing the 87 core t(m)DNA types typically
415 present in *Klebsiella* spp. genomes are shown using *K. pneumoniae* MGH78578 as the reference
416 strain. For each locus, the conserved flanking gene context is displayed. Scale bars may vary
417 between panels to optimize visualization. Gene names are indicated when available; otherwise,
418 locus tags are provided.

419 **Figure S2. Strategy for genomic island identification based on conserved genomic context**
420 **and manual curation.** Representative examples of integration site identification in two strains
421 from the GMI-K70 collection are shown. In *K. pneumoniae* HS11286, the conserved genomic
422 context surrounding the *pro2A* tDNA is preserved. In contrast, in strain PMK1, disruption of this
423 conserved architecture indicates insertion of a genomic island. Following detection of this synteny
424 shift, the genomic island was manually delineated. The integrase gene encoded within the MGE
425 (int, shown in red) and the flanking direct repeats (gray boxes) were subsequently identified. For

426 integration events occurring within protein-coding genes or intergenic regions, the same principle
427 was applied, with integrase detection serving as an initial indicator of candidate integration sites.

428 **Table S1. Metadata of strains included in the GMI-K70 collection.** Summary of species
429 assignments, sequence types (STs), year of isolation, country of origin, and genome accession
430 numbers for the 70 genomes included in the GMI-K70 collection. (provided as a separate
431 spreadsheet).

432 **Table S2. Catalogue and prevalence of identified t(m)DNA types.** Summary of the 157
433 t(m)DNA types identified across the GMI-K70 collection, including their prevalence among strains
434 and nucleotide sequence identity statistics. For each type, all pairwise nucleotide identities were
435 calculated and averaged. The minimum observed pairwise identity is also provided. (provided as
436 a separate spreadsheet).

437
438 **Table S3. Chromosomal integration sites and associated mobile genetic elements.** List of
439 the 67 chromosomal integration loci identified in the GMI-K70 collection, comprising 21 t(m)DNA
440 loci and 46 non-t(m)DNA loci. For each site, integration frequency and representative associated
441 MGEs are provided. (provided as a separate spreadsheet).

442 **Table S4. Structural and compositional characteristics of curated MGEs.** Summary of the
443 676 chromosomal MGEs identified in the GMI-K70 collection, including length, GC content,
444 presence or absence of integrase genes, and detection of flanking direct repeats (DRs). (provided
445 as a separate spreadsheet).

446 **AUTHOR CONTRIBUTIONS**

447 Camilo Berríos-Pastén: Conceptualization, Data curation, Formal analysis, Investigation,
448 Methodology, Software, Visualization, Writing – original draft, Writing – review and editing.
449 Rodolfo Acevedo: Data curation, Formal analysis, Investigation, Software. Carlos Serrano-Pinto:
450 Data curation, Investigation, Methodology. Patricio Arros: Investigation, Validation, Methodology,
451 Writing – review and editing. Margaret M. C. Lam: Methodology, Validation, Writing – review and
452 editing. Kathryn E. Holt: Conceptualization, Resources, Supervision, Writing – review and editing.
453 Rosalba Lagos: Conceptualization, Supervision, Writing – review and editing, Funding acquisition.
454 Kelly L. Wyres: Conceptualization, Resources, Supervision, Writing – review and editing. Andrés
455 E. Marcoleta: Conceptualization, Project administration, Resources, Supervision, Visualization,
456 Writing – original draft, Writing – review and editing, Funding acquisition.

457 **CONFLICTS OF INTEREST**

458 The author(s) declare that there are no conflicts of interest.

459 **FUNDING INFORMATION**

460 This study was funded by the National Agency for Research and Development (ANID, Chile),
461 Grant FONDECYT 1221193. Camilo Berríos-Pastén was funded by the scholarship
462 Becas/Doctorado Nacional 2019-21192024.

463 **ETHICAL APPROVAL**

464 Not applicable.

465 **CONSENT FOR PUBLICATION**

466 Not applicable.

467 **REFERENCES**

- 468 1. Fang CT, Lai SY, Yi WC, Hsueh PR, Liu KL, Chang SC. *Klebsiella pneumoniae* Genotype
469 K1: An Emerging Pathogen That Causes Septic Ocular or Central Nervous System
470 Complications from Pyogenic Liver Abscess. *Clinical Infectious Diseases*.
471 2007;45(3):284–93. doi:10.1086/519262 PubMed PMID: 17599305.
- 472 2. Siu LK, Yeh KM, Lin JC, Fung CP, Chang FY. *Klebsiella pneumoniae* liver abscess: A new
473 invasive syndrome. *Lancet Infect Dis*. 2012;12(11):881–5. doi:10.1016/S1473-
474 3099(12)70205-0 PubMed PMID: 23099082.
- 475 3. Chung The H, Karkey A, Pham Thanh D, Boinett CJ, Cain AK, Ellington M, et al. A high-
476 resolution genomic analysis of multidrug-resistant hospital outbreaks of *Klebsiella*
477 *pneumoniae*. *EMBO Mol Med*. 2015;7(3):227–39. doi:10.15252/emmm.201404767
478 PubMed PMID: 25712531.
- 479 4. Wyres KL, Holt KE. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes
480 from environmental to clinically important bacteria. *Curr Opin Microbiol*. 2018;45:131–9.
481 doi:10.1016/j.mib.2018.04.004 PubMed PMID: 29723841.
- 482 5. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic
483 analysis of diversity, population structure, virulence, and antimicrobial resistance in
484 *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A*.
485 2015;112(27):E3574–81. doi:10.1073/pnas.1501049112 PubMed PMID: 26100894.
- 486 6. Marcoleta AE, Berríos-Pastén C, Nuñez G, Monasterio O, Lagos R. *Klebsiella*
487 *pneumoniae* asparagine tDNAs are integration hotspots for different genomic Islands
488 encoding microcin E492 production determinants and other putative virulence factors

- 489 present in hypervirulent strains. *Front Microbiol.* 2016;7(JUN).
490 doi:10.3389/fmicb.2016.00849
- 491 7. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic
492 diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in
493 *Klebsiella pneumoniae* populations. *Microb Genom.* 2018;1–14.
494 doi:10.1099/mgen.0.000196 PubMed PMID: 26345333.
- 495 8. Villa L, Feudi C, Fortini D, Brisse S, Passet V, Bonura C, et al. Diversity, virulence, and
496 antimicrobial resistance of the KPC-producing *Klebsiella pneumoniae* ST307 clone.
497 *Microb Genom.* 2017;3(4). doi:10.1099/mgen.0.000110 PubMed PMID: 28785421.
- 498 9. Juhas M, Van Der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic
499 islands: Tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev.*
500 2009;33(2):376–93. doi:10.1111/j.1574-6976.2008.00136.x PubMed PMID: 19178566.
- 501 10. Gal-Mor O, Finlay BB. Pathogenicity islands: A molecular toolbox for bacterial virulence.
502 *Cell Microbiol.* 2006;8(11):1707–19. doi:10.1111/j.1462-5822.2006.00794.x PubMed
503 PMID: 16939533.
- 504 11. Keiler KC. Biology of *trans*-Translation. *Annu Rev Microbiol.* 2008;62(1):133–51.
505 doi:10.1146/annurev.micro.62.081307.162948 PubMed PMID: 18557701.
- 506 12. Bellanger X, Payot S, Leblond-Bourget N, Guédon G. Conjugative and mobilizable
507 genomic islands in bacteria: Evolution and diversity. *FEMS Microbiol Rev.*
508 2014;38(4):720–60. doi:10.1111/1574-6976.12058 PubMed PMID: 24372381.
- 509 13. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes:
510 sublocation preference of integrase subfamilies. *Nucleic Acids Res.* 2002;30(4):866–75.
511 doi:10.1093/nar/30.4.866 PubMed PMID: 11842097.
- 512 14. Tan YH, Arros P, Berríos-Pastén C, Wijaya I, Chu WHW, Chen Y, et al. Hypervirulent
513 *Klebsiella pneumoniae* employs genomic island encoded toxins against bacterial
514 competitors in the gut. *ISME Journal.* 2024;18(1). doi:10.1093/ismejo/wrae054 PubMed
515 PMID: 38547398.
- 516 15. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the
517 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
518 *Genome Res.* 2015;25(7):1043–55. doi:10.1101/gr.186072.114 PubMed PMID:
519 25977477.
- 520 16. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. A genomic surveillance
521 framework and genotyping tool for *Klebsiella pneumoniae* and its related species
522 complex. *Nat Commun.* 2021;12(1). doi:10.1038/s41467-021-24448-3 PubMed PMID:
523 34234121.
- 524 17. Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, Criscuolo A, et al. A Dual
525 Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of *Klebsiella*
526 *pneumoniae* Strains. *Mol Biol Evol.* 2022;39(7). doi:10.1093/molbev/msac135 PubMed
527 PMID: 35700230.

- 528 18. Argimón S, David S, Underwood A, Abrudan M, Wheeler NE, Kekre M, et al. Rapid
529 Genomic Characterization and Global Surveillance of *Klebsiella* Using Pathogenwatch.
530 Clinical Infectious Diseases. 2021 Dec 1;73(Supplement_4):S325–35.
531 doi:10.1093/cid/ciab784
- 532 19. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing
533 polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21(1):1–
534 21. doi:10.1186/s13059-020-02090-4 PubMed PMID: 32698896.
- 535 20. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et
536 al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
537 Genomic Era. Mol Biol Evol. 2020;37(5):1530–4. doi:10.1093/molbev/msaa015 PubMed
538 PMID: 32011700.
- 539 21. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in
540 nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–6. doi:10.1093/nar/gkh152
541 PubMed PMID: 14704338.
- 542 22. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of
543 transfer RNA genes. Nucleic Acids Res. 2016;44(W1):W54–7. doi:10.1093/nar/gkw413
544 PubMed PMID: 27174935.
- 545 23. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta:
546 Rapid and standardized annotation of bacterial genomes via alignment-free sequence
547 identification. Microb Genom. 2021;7(11). doi:10.1099/MGEN.0.000685 PubMed PMID:
548 34739369.
- 549 24. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and
550 environmental microorganisms. Nat Rev Microbiol. 2004 May 1;2(5):414–24.
551 doi:10.1038/nrmicro884 PubMed PMID: 15100694.
- 552 25. Dobrindt U, Reidl J. Pathogenicity islands and phage conversion: Evolutionary aspects of
553 bacterial pathogenesis. International Journal of Medical Microbiology. 2000;290(6):519–
554 27. doi:10.1016/S1438-4221(00)80017-X PubMed PMID: 11100826.
- 555 26. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and
556 environmental microorganisms. Nat Rev Microbiol. 2004;2(5):414–24.
557 doi:10.1038/nrmicro884 PubMed PMID: 15100694.
- 558 27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J
559 Mol Biol. 1990 Oct;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2
- 560 28. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of
561 progressive multiple sequence alignment through sequence weighting, position-specific
562 gap penalties and weight matrix choice. Nucleic Acids Res. 1994 Nov 11;22(22):4673–80.
563 PubMed PMID: 7984417.
- 564 29. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the
565 analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8. doi:10.1038/nbt.3988
566 PubMed PMID: 29035372.

- 567 30. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: A comparative pathogenomic
568 platform with an interactive web interface. *Nucleic Acids Res.* 2019;47(D1):D687–92.
569 doi:10.1093/nar/gky1080
- 570 31. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al.
571 AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic
572 links among antimicrobial resistance, stress response, and virulence. *Sci Rep.*
573 2021;11(1):1–9. doi:10.1038/s41598-021-91456-0 PubMed PMID: 34135355.
- 574 32. Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, et al. PHASTEST: Faster than
575 PHASTER, better than PHAST. *Nucleic Acids Res.* 2023;51(W1):W443–50.
576 doi:10.1093/nar/gkad382 PubMed PMID: 37194694.
- 577 33. Piña-Iturbe A, Hoppe-Elsholz G, Suazo ID, Kalergis AM, Bueno SM. Subinhibitory
578 antibiotic concentrations promote the excision of a genomic island carried by the globally
579 spread carbapenem-resistant *Klebsiella pneumoniae* sequence type 258. *Microb Genom.*
580 2023 Dec 11;9(12). doi:10.1099/mgen.0.001138
- 581