

Research article

1 **Title**

2 Genomic epidemiology and global diversity of the emerging bacterial pathogen

3 *Elizabethkingia anophelis*

4

5 **Authors and affiliations**

6 Sebastien Breurec^{a,b,#}, Alexis Criscuolo^c, Laure Diancourt^d, Olaya Rendueles^{e,f}, Mathias

7 Vandenbergert^d, Virginie Passet^{e,f}, Valérie Caro^d, Eduardo P. C. Rocha^{e,f}, Marie Touchon^{e,f},

8 Sylvain Brisse^{e,#}

9

10 Unité Environnement et Santé, Institut Pasteur de Guadeloupe, Pointe-à-Pitre, France, Faculté

11 de Médecine Hyacinthe Bastaraud, Université des Antilles, Pointe-à-Pitre, France,

12 Laboratoire de Microbiologie clinique et environnementale, Centre Hospitalier Universitaire

13 de Pointe-à-Pitre/les Abymes, Pointe-à-Pitre, France^a; Laboratoire de Microbiologie

14 Clinique, Institut Pasteur de Bangui, Bangui, Central African Republic^b; Institut Pasteur, Hub

15 Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, Paris, France^e; Unité

16 Environnement et Risques Infectieux, Institut Pasteur, Paris, France^d; Microbial Evolutionary

17 Genomics, Institut Pasteur, Paris, France^e; UMR 3525, CNRS, Paris, France^f

18

19 **Running head**

20 Genomic diversity of *Elizabethkingia anophelis*

21

22 # Address correspondence to Sebastien Breurec, sbreurec@gmail.com, and Sylvain Brisse,

23 sbrisse@pasteur.fr

24 S. Breurec and AC contributed equally. MT and S. Brisse co-supervised the work.

25

26

27 **Word count**

28 Abstract: 244, text: 5556 words

29

30

31 **ABSTRACT**

32 *Elizabethkingia anophelis* is an emerging pathogen. Genomic analysis of strains from
33 clinical, environmental or mosquito sources is needed to understand the epidemiological
34 emergence of *E. anophelis* and to uncover genetic elements implicated in antimicrobial
35 resistance, pathogenesis, or niche adaptation. Here, the genomic sequences of two
36 nosocomial isolates that caused neonatal meningitis in Bangui, Central African Republic,
37 were determined and compared with *Elizabethkingia* isolates from other world regions and
38 sources. Average nucleotide identity firmly confirmed that *E. anophelis*, *E. meningoseptica*
39 and *E. miricola* represent distinct genomic species and led to re-identification of several
40 strains. Phylogenetic analysis of *E. anophelis* strains revealed several sublineages and
41 demonstrated a single evolutionary origin of African clinical isolates, which carry unique
42 antimicrobial resistance genes acquired by horizontal transfer. The *Elizabethkingia* genus and
43 the species *E. anophelis* had pan-genomes comprising respectively 7,801 and 6,880 gene
44 families, underlining their genomic heterogeneity. African isolates were capsulated and
45 carried a distinctive capsular polysaccharide synthesis cluster. A core-genome multilocus
46 sequence typing scheme applicable to all *Elizabethkingia* isolates was developed, made
47 publicly available (<http://bigsd.dbweb.pasteur.fr/elizabethkingia>), and shown to provide useful
48 insights into *E. anophelis* epidemiology. Furthermore, a clustered regularly interspaced short
49 palindromic repeats (CRISPR) locus was uncovered in *E. meningoseptica*, *E. miricola* and in
50 a few *E. anophelis* strains. CRISPR spacer variation was observed between the African
51 isolates, illustrating the value of CRISPR for strain subtyping. This work demonstrates the
52 dynamic evolution of *E. anophelis* genomes and provides innovative tools for *Elizabethkingia*
53 identification, population biology and epidemiology.

54

55

56

57 **IMPORTANCE**

58 *Elizabethkingia anophelis* is a recently recognized bacterial species involved in human
59 infections and outbreaks in distinct world regions. Using whole-genome sequencing, we
60 showed that the species comprises several sublineages, which differ markedly in their
61 genomic features associated with antibiotic resistance and host-pathogen interactions.
62 Further, we have devised high-resolution strain subtyping strategies and provide an open
63 genomic sequence analysis tool, facilitating the investigation of outbreaks and tracking of
64 strains across time and space. We illustrate the power of these tools by showing that two
65 African healthcare-associated meningitis cases observed 5 years apart were caused by the
66 same strain, providing evidence that *E. anophelis* can persist in the hospital environment.

67 **INTRODUCTION**

68 *Elizabethkingia* is a genus of aerobic, non-motile, Gram-negative rods belonging to
69 family *Flavobacteriaceae* in the phylum *Bacteroidetes*. *Elizabethkingia* strains are considered
70 ubiquitous in nature, and the genus currently comprises three species. *E. meningoseptica* is a
71 common cause of neonatal meningitis, especially in premature infants during the first weeks
72 of life, and also causes a wide range of human infections as an opportunistic pathogen (1-3).
73 *E. miricola* is rarely associated with sepsis (4). *E. anophelis* was recently described from the
74 midgut of *Anophelis gambiae* mosquitoes (5). Shortly after, some of us described the two first
75 human cases of neonatal meningitis, caused by *E. anophelis* in Bangui, Central African
76 Republic (6). A nosocomial origin was suspected for these strains given that patients had a
77 history of intubation and mechanical ventilation at birth. The two meningitis cases occurred
78 in the same hospital five years apart and an epidemiological link between them was not
79 established. *E. anophelis* was also implicated in an outbreak in an intensive care unit in
80 Singapore (7) and in a transmission from mother to infant in Hong Kong (8), resulting in late-
81 onset meningitis in the baby. At the time of writing, an unprecedented large outbreak of
82 *E. anophelis* from a yet unknown source is ongoing in Wisconsin, USA. *E. anophelis* is
83 therefore considered as an emerging opportunistic pathogen. In all previously reported
84 clinical cases (6-8), *E. anophelis* isolates were resistant to most antibiotics recommended for
85 empirical therapy, including third-generation cephalosporins, aminoglycosides and
86 carbapenems (9), probably accounting for the poor infection outcomes. Consistently, most
87 *E. anophelis* genomes harbor multiple antimicrobial resistance genes, including the two
88 metallo-beta-lactamase genes *bla_{GOB}* and *bla_{blaB}*, associated with resistance to carbapenems
89 (10). Currently, the clinical significance and biological distribution of *E. anophelis* is
90 restricted, as this novel species is not well distinguished from other *Elizabethkingia* species in

91 routine microbiological identification procedures, and as reference identification of
92 *E. anophelis* relies mostly on the low-resolution 16S rRNA gene.

93 Genomic studies of *E. anophelis* strains from clinical, environmental and mosquito
94 origins have revealed interesting features that may contribute to the adaptation of this
95 bacterium to mosquito and that may be implicated in its pathogenesis (8, 11, 12). Genomic
96 sequencing has also been used to derive epidemiological insights during hospital outbreaks
97 (8, 12). However, knowledge on the genomic diversity and population structure of
98 *E. anophelis* is still scant, and there is currently no standard way to compare isolates at local,
99 regional or global scales, which limits our understanding of the epidemiology and global
100 emergence of *E. anophelis*.

101 The aims of this study were to determine the phylogenetic origin and genomic
102 features of the two African neonatal meningitis *E. anophelis* isolates; to compare them with
103 *Elizabethkingia* isolates from other world regions and sources; to take advantage of the high
104 resolution of whole genome sequencing to distinguish *Elizabethkingia* species
105 unambiguously, and to develop a publicly-accessible genotyping system enabling a common
106 language on strain types in order to ease future collaborations on biodiversity studies and
107 epidemiological investigations of *Elizabethkingia* clinical and environmental isolates.

108 **RESULTS AND DISCUSSION**

109

110 **Genome sequences of African isolates E27017 and E18064**

111 The two African isolates were sequenced using paired-end Illumina technology. The
112 genome sequence of isolate E18064 was assembled into 213 contigs of total length 4,080,763
113 bp. It is 35.7% G+C rich and contains 3,648 CDSs. The E27107 genome sequence is
114 4,059,474 bp long, 35.5% G+C rich and contains 3,674 CDSs distributed over 89 contigs.
115 The genome characteristics of the two African isolates thus appeared similar to those of
116 previously sequenced *Elizabethkingia* genomes (Table S1).

117

118 **Genome-based identification of members of the *Elizabethkingia* genus**

119 To determine with confidence the species-level identification of the two African and
120 the 18 other *Elizabethkingia* isolates, we calculated average nucleotide identity (ANI) based
121 on their genomic sequences (Table S2). The distribution of ANI values among the 20
122 genomes classified them into three main groups (Fig. 1a) and showed a clear separation of
123 *E. anophelis* strains from *E. meningoseptica* and *E. miricola*. These results fully demonstrate
124 that these three taxa represent distinct genomic groups. The two African isolates clearly
125 belong to *E. anophelis* based on their high ANI with R26^T, the taxonomic type strain of this
126 species, fully confirming their initial identification (6). Even though they were initially
127 identified as *E. meningoseptica*, the three strains 502, Endophthalmitis and B2D also
128 appeared closely related to *E. anophelis* isolates based on ANI values. Therefore, these three
129 strains should be re-assigned to *E. anophelis*. Based on its ANI value with *E. miricola*
130 reference strain ATCC 33958, the unclassified *Elizabethkingia* sp. strain BM10 isolated from
131 a termite, clearly belongs to *E. miricola*. Urea degradation is a differential phenotypic
132 characteristic of *E. miricola*, and we confirmed the presence in both *E. miricola* genomes of

133 the urease cluster, whereas it was absent from all other genomes. The ANI values among
134 *E. anophelis* genomes ranged from 96.97% to 100% (Fig. 1b), thus clearly higher than the
135 classical 95% threshold used for species delineation. This result confirms that *E. anophelis*
136 isolates belong to a single genomic species. Of note, the ANI value estimated between
137 *E. anophelis* and *E. meningoseptica* was only 79.7%, showing that these species are separated
138 by an evolutionary distance similar to the one observed between *Escherichia coli* and
139 *Salmonella* spp. (80.4%), two model species with very distinct ecological lifestyles and
140 virulence features.

141 The two African *E. anophelis* isolates were initially identified based on 16S rRNA
142 sequencing (6). Phylogenetic analysis of the 16S rRNA sequences (Fig. S1a) was consistent
143 with the clustering into three species. However, four *E. anophelis* isolates (502, B2D,
144 Endophthalmitis and PW2810) were linked to the other *E. anophelis* strains in a clade with an
145 atypically long branch. Inspection of the alignment revealed that this long branch was caused
146 by 10 nucleotide differences clustered within a single region of 35 nucleotides (Fig. S1b).
147 These results indicate that careful interpretation may be required when classifying
148 *Elizabethkingia* isolates based on the divergence values estimated from the classical 16S
149 rRNA marker.

150

151 **Gene repertoire of *Elizabethkingia***

152 To compare the gene repertoire of the two African isolates with other *Elizabethkingia*
153 isolates, we first quantified the diversity of *Elizabethkingia* genes by computing the set of
154 ubiquitous genes (core genome) and the set of different homologous genes families (pan-
155 genome) among the 20 genomes. At the level of the genus *Elizabethkingia*, the core genome
156 contained 2,221 orthologous protein families, corresponding to 65% of the size of the
157 smallest proteome observed for *E. meningoseptica* NBRC 12535. The core-genome of the

158 species *E. anophelis* had 2,512 orthologous protein families, corresponding to 72% of the size
159 of the smallest proteome (strain B2D). Gene rarefaction analyses showed that the
160 *E. anophelis* core genome varies little with the addition of the last genomes (Fig. 2a),
161 suggesting that this core genome estimate is robust. Both the pan-genomes of the
162 *Elizabethkingia* genus and of *E. anophelis* were large with, respectively, 7,801 and 6,880
163 gene families. Gene rarefaction analyses showed that the addition of future *E. anophelis*
164 genomes to the analysis will still significantly increase the size of the pan-genome (12). The
165 spectrum of gene frequencies for the *E. anophelis* pan-genome (Fig. 2b) showed that the vast
166 majority of gene families were either encoded in a few genomes (39% in three or less) or in
167 most of them (42% in more than 14 genomes), confirming that the genome sequencing of
168 additional *Elizabethkingia* strains will uncover multiple novel gene families.

169

170 **Core genome-based phylogenetic structure of *E. anophelis***

171 To determine the phylogenetic origin of the African strains with respect to strains
172 from other world regions, we selected a subset of 1,546 genes families with very reliable
173 alignments (few indels) to infer robust phylogenies. Phylogenetic analysis based on
174 concatenated alignments of the 1,546 gene sequences showed that all *E. anophelis* isolates
175 were clearly separated from closely related species *E. miricola* (its sister group) and
176 *E. meningoseptica* (Fig. 3a). These results fully corroborated the classification into three
177 genomic groups based on ANI and 16S rRNA analyses.

178 The phylogenetic structure of *E. anophelis* revealed two main lineages, each
179 containing distantly related sublineages (Fig. 3b). Whereas clinical isolates PW2810, 502,
180 B2D and Endophthalmitis formed lineage B, all other human infection isolates were clustered
181 into lineage A together with the mosquito strains Ag1 and R26. Therefore, *E. anophelis*
182 mosquito strains share a common evolutionary history with a subset of *Elizabethkingia*

183 isolates from humans. Nevertheless, the two mosquito strains were closely related and
184 represented a distinct sublineage (named sublineage 1) that was separated by a large
185 evolutionary distance from other isolates. Four additional sub-lineages were discerned within
186 major lineage A, comprising respectively: i) Three patient isolates and two hospital
187 environment isolates from a Singapore outbreak (7); ii) Two unrelated hospital environment
188 isolates from the above outbreak investigation; iii) Strains PW2806 and PW2809 (HKU37
189 and HKU38, respectively) from Hong Kong (8); and iv) The two African strains from Bangui
190 (6). Within each sublineage, the strains were highly related. These results show that the
191 African strains formed a distinct sublineage, therefore having a clearly distinct evolutionary
192 origin from Asian nosocomial strains and mosquito strains. The nearly simultaneous
193 recognition of the emergence of *E. anophelis* infections in Africa and Asia thus reflects
194 independent epidemiological events.

195

196 **Antimicrobial resistance-associated genomic features of African *E. anophelis* isolates**

197 All *Elizabethkingia* genomes contained at least 17 antimicrobial resistance genes
198 (Table S3). Most prominent among these were genes coding for beta-lactamases, including
199 the metallo-beta-lactamase genes *bla_{blaB}* and *bla_{GOB}* (10, 13), and for efflux systems, which
200 may contribute largely to the phenotypic resistance of *Elizabethkingia* isolates to most
201 antibiotics. These findings are consistent with previous descriptions of multiple resistance
202 genes in *Elizabethkingia* genomes (8, 11, 12, 14). Most resistance genes belonged to
203 persistent families, i.e., they were part of the core genome of *Elizabethkingia*. This was the
204 case for genes *bla_{blaB}* and *bla_{GOB}*, which were previously described as intrinsic and
205 chromosome-borne in *E. meningoseptica* (10). We observed that both *bla_{blaB}* and *bla_{GOB}* were
206 located in highly syntenic regions in all members of the *Elizabethkingia* genus, but were
207 absent from the closely related genera *Flavobacterium*, *Riemerella*, *Chryseobacterium* (not

208 shown). These results show that multiple potential resistance genes are ancestral in the genus
209 *Elizabethkingia* and suggest acquisition by horizontal gene transfer into this genus before the
210 evolutionary separation of the three *Elizabethkingia* species.

211 Interestingly, several additional resistance genes were detected in the African isolates
212 (24 genes in total) as compared to other *Elizabethkingia* genomes (range: 17 – 20). Whereas
213 all *Elizabethkingia* genomes contained one conserved chloramphenicol acetyltransferase
214 (CAT) gene, the two African isolates contained an additional CAT gene that had 80.7%
215 amino-acid identity with the conserved copy. The phylogenetic relationships of the conserved
216 copy (Fig. S2) was concordant with the classification of the strains into *E. meningoseptica*,
217 *E. miricola* and *E. anophelis*, indicating vertical evolution of the conserved CAT copy from
218 an ancestral *Elizabethkingia* gene. In contrast, the additional CAT copy from African isolates
219 was closely related to *Riemerella* sequences (Fig. S2), suggesting introduction by horizontal
220 gene transfer from this genus.

221 Each of the African isolates also carried two additional metallo-beta-lactamases of the
222 B1 (NDM-CcrA) subclass. In addition, an aminoglycoside acetyltransferase (AAC3-I) was
223 found only in the two African isolates and in one isolate (NUH6) from Singapore. Finally, a
224 *tet(X)* gene coding for a tetracycline inactivating enzyme 99.7% identical to that of
225 *Bacteroides fragilis* (15) was found in the African isolates and in the Singapore isolate
226 NUH4. Interestingly, the additional CAT gene, *tet(X)* and one subclass B1 beta-lactamase
227 gene were clustered in a genomic region that also contained a putative class D beta-lactamase
228 gene, a mercuric reductase gene and a lincosamide nucleotidyltransferase (LinF) gene
229 associated with resistance to lincosamides (16). This region had high similarity to plasmid
230 pRA0511 of *Riemella anatipestifer* (17). Furthermore, the other subclass B1 beta-lactamase
231 gene and the AAC3-I gene were adjacent to each other and to a transposase gene. These

232 observations indicate that the additional resistance genes of the African *E. anophelis* isolates
233 were acquired through the transfer of mobile genetic elements.

234

235 **Identification and diversity of capsular polysaccharide synthesis clusters**

236 Currently, the pathophysiology of *Elizabethkingia* infections is poorly understood (18).
237 We searched for virulence-associated features in the African *E. anophelis* and other
238 *Elizabethkingia* genomes by using the VFDB database (Table S4). A number of coding
239 sequences showed homologs to iron and heme acquisition, to hemolysins and to oxidative
240 stress resistance proteins, consistent with previous reports (8, 11, 12, 14). Among the hits
241 found in the VFDB database, several corresponded to proteins putatively involved in capsule
242 synthesis.

243 To our knowledge, *Elizabethkingia* strains were not previously reported as being
244 capsulated or as harboring a *cps* cluster. Given that bacterial capsules represent one of the
245 main bacterial virulence factors, particularly in pathogens that may be responsible for
246 meningitis (19), we investigated whether the African isolates are capsulated. Using the India
247 ink method, we observed a clearly distinguishable capsule, produced by the two strains
248 during stationary phase when cultured either in BHI or in LB medium.

249 Further inspection of genome sequences revealed a Wzy-dependent *cps* gene cluster
250 (20) comprising 27 co-oriented coding sequences in the two African isolates and in most
251 other *Elizabethkingia* strains (Fig. 4). The upstream (3') part of the gene cluster was highly
252 conserved (Fig. 4) and included the two well-characterized genes *wzc* and *wza*. The former
253 codes for an inner membrane protein necessary for capsular polymerization and translocation
254 across the inner membrane. The latter encodes a channel that allows translocation of the
255 polysaccharide across the outer membrane. Interestingly, these two genes were found in two

256 copies in most strains separated by three other conserved genes (a RecX family transcription
257 factor, and two other genes of unknown function).

258 The downstream region was highly variable and may be responsible for the Wzy-
259 dependent synthesis of polysaccharides with distinctive chemical compositions and which
260 would possibly correspond to distinct capsular serotypes as observed in other capsulated
261 species (21, 22). This region comprised the flippase (*wzx*) and the polymerase (*wzy*) and other
262 genes coding for glycosyl transferases, acetyl transferases and other sugar-modifying
263 enzymes. In some strains, the essential genes involved in polymerization - *wzx* and *wzy* -
264 could not be confidently identified. These genes are known to be poorly conserved even
265 among phylogenetically closely related isolates, rendering their detection difficult (23).

266 Interestingly, the genetic organization was strongly conserved among strains belonging
267 to the same phylogenetic sublineage (as defined above; Fig. 4). These results suggest that
268 capsular types are conserved within *E. anophelis* sublineages, but not between them. For
269 example, the groups of African isolates, and mosquito isolates (R26 and Ag1) showed few
270 intra-group and many between-group differences in the downstream regions of the capsular
271 cluster locus. More specifically, the two African genomes seem to code for a unique and
272 distinctive capsule, relative to other *Elizabethkingia* strains. Capsular type variation in
273 *Elizabethkingia* may restrict immune cross-reactions and could be associated with distinct
274 pathogenic properties, as observed in other capsulated bacterial groups (24-26). Overall,
275 these results warrant future studies into the pathophysiological and epidemiological
276 implications of capsular variation in *Elizabethkingia*.

277

278 **Secretion systems**

279 Bacterial secretion systems are associated with important functions such as detoxification,
280 antibiotic resistance and scavenging. We used MacSyFinder (27), together with TXSScan

281 profiles (28), to identify the secretion systems encoded in the *Elizabethkingia* genomes (see
282 Methods and **Table S5**). While T9SS (PorSS) is present in many members of the phylum
283 *Bacteroidetes* (29), we found none in the *Elizabethkingia* genomes. We also found no
284 evidence for the presence of Type 2, 3, 5 protein secretion systems, which fits previous
285 suggestions that these systems are absent from *Bacteroidetes* (28). In contrast, we found two
286 families of T1SS in the genomes of *Elizabethkingia*, one of which was present in both
287 African strains.

288 A third variant of the type VI secretion system (T6SSiii) was recently uncovered in
289 *Bacteroidetes* (30). T6SS can be involved in bacterial competition (30) or pathogenicity (31).
290 Interestingly, one locus encoding a T6SSⁱⁱⁱ was found in every strain of our dataset. This
291 system was highly conserved, except for the gene family *tssI* that showed extensive copy
292 number variability. This gene encodes the cell-puncturing device and it can be fused with a
293 variety of toxic domains, thus encoding a toxin (32). These systems might provide
294 *Elizabethkingia* strains a way to antagonize competing bacteria in the complex environments
295 such as the mosquito gut.

296 Type 4 secretion systems (T4SS) implicated in conjugative DNA transfer were
297 previously described in *E. anophelis* (11, 14). We identified such systems in nearly all the
298 genomes. Some strains encoded up to six conjugative systems (e.g., strain NUHP3), which is
299 among the highest values observed among bacteria (33). Expectedly, all these systems were
300 of the MPFB type, the one specific to *Bacteroidetes*. All the genomes encoding a T4SS also
301 encoded one or several relaxases, suggesting that these systems are involved in conjugation
302 of mobile genetic elements and not in protein secretion. The abundance of conjugative mobile
303 genetic elements suggests that they might drive much of the genetic diversification observed
304 in *Elizabethkingia*.

305

306 **Standardized strain typing method based on genome sequencing for *Elizabethkingia***
307 **epidemiology**

308 In order to explore the epidemiological links between the African and other
309 *E. anophelis* isolates, we translated whole genome sequences into high-resolution genotyping
310 data. For this purpose, we followed the cgMLST strategy, which allows creating universal
311 genotype nomenclatures and comparing isolates globally (34-36). A cgMLST scheme was
312 developed based on 1,542 genes that could be aligned with high confidence among
313 *Elizabethkingia* genomes. We then tested this scheme by scanning the 20 isolates for allelic
314 variation at these loci. The identity of the 1,542 alleles was recorded for each isolate. The
315 resulting allelic profiles were then clustered to obtain groups of isolates. These groups
316 coincided with the sublineages from the phylogenetic analysis of the core gene sequences
317 (Fig. S4).

318 To validate our cgMLST approach, we took advantage of the availability within the
319 dataset of public genomes, of groups of isolates that were epidemiologically related, i.e.,
320 which were isolated during single outbreaks or were associated with direct mother-to-child
321 transmission (7, 8). First, the five *E. anophelis* strains NUHP1, NUHP2, NUHP3, NUH1 and
322 NUH4 from a single outbreak from Singapore were compared (7). We found that four of
323 these strains were almost identical (2 to 6 cgMLST allelic mismatches), whereas NUH4 was
324 slightly more distant (32 or more allelic mismatches with the others). This is highly
325 consistent with the whole genome single nucleotide polymorphism (SNP) analysis by Teo
326 and colleagues (12), who found 24 to 38 SNPs among the four closely related isolates,
327 whereas at least 176 SNPs separated these from isolate NUH4. In a second study (8), two
328 isolates (HKU37 and HKU38) from a mother and her neonate were considered to be directly
329 related based on their genomic sequences, which showed no difference out of 2,000 genes.
330 Based on our cgMLST approach, these two isolates were also identical, as none of the 1,542

331 loci was variable between them. In contrast, 1,433 cgMLST loci were distinct between this
332 pair of isolates and a third independent isolate (HKU36) (8). These results demonstrate the
333 extremely high efficiency of the cgMLST approach to distinguish epidemiologically related
334 isolates from unrelated ones.

335 To our surprise, the two African *E. anophelis* isolates, which were isolated 5 years
336 apart in the same hospital in Bangui (6), differed at only 4 alleles. This level of variation is
337 similar to the variation observed among related isolates of the Singapore outbreak and thus
338 strongly suggests to us that the two African hospital isolates may also be epidemiologically
339 related. We therefore hypothesize that a single *E. anophelis* strain caused the two infections
340 five years apart, due to its persistence in a hospital environmental source.

341 The core-genome MLST approach and attached genomic sequences were made
342 publicly available through the Institut Pasteur server
343 (<http://bigsdbs.web.pasteur.fr/elizabethkingia/>).

344

345 **Clustered regularly interspaced short palindromic repeats (CRISPR)**

346 CRISPR loci exhibit spacer diversity that reflects a history of prior invasions by
347 different phages and plasmids (37). In addition, the high variability of CRISPR loci in some
348 bacterial species implies that they represent useful markers for epidemiological investigations
349 (38, 39). To extend our understanding of genomic variation between the African isolates and
350 among the other *Elizabethkingia* isolates, we investigated CRISPR-Cas systems and analyzed
351 spacer content in the 20 strains. All three species contained subtype II-C CRISPR-Cas
352 systems comprising of Cas1, Cas2, and Cas9 proteins, while no homolog of Cas4 and Csn2
353 were identified (40). The loci were characterized by unusually long repeats (that is, 47 bp in
354 size, Fig. 5a), as previously observed in bacteria belonging to the *Bacteroidetes* phylum (40).
355 However, the CRISPR locus was only present in a few *Elizabethkingia* strains (Fig. S3).

356 These results demonstrate the occurrence of CRISPR systems in *Elizabethkingia* and the
357 variable presence of these systems among phylogenetic sublineages of *E. anophelis*.

358 Analysis of the spacer diversity among *Elizabethkingia* CRISPR arrays revealed no
359 shared spacer between or within species, with the exception of the two hospital acquired
360 African isolates, which showed a strong CRISPR conservation (Fig. 5b). This finding
361 suggests a high rate of gain and loss of complete CRISPR-Cas system during *E. anophelis*
362 evolution. CRISPR arrays of the two African isolates were almost identical, with the
363 exception that E18064, which was isolated 5 years later (in 2011), contained two additional
364 spacers at the 5' end of the array (called S22 and S23, Fig. 5b). Importantly, according to the
365 current spacer acquisition model, these two spacers are the most recent acquired spacers (see
366 Methods). This clear genetic difference between CRISPR arrays of the African isolates was
367 confirmed by PCR. CRISPR-based typing may thus offer the high level of discrimination
368 needed for strain subtyping in a local epidemiological context.

369 Surprisingly, while the S22 and S23 spacers were clearly absent in the CRISPR array
370 of the earlier African isolate E27107 (isolated in 2006), we found perfect matches (*i.e.*
371 protospacers) in one E27107 specific genomic region of 43 kb (not including the CRISPR
372 array) (Fig. 5c). This particular genomic island was totally absent in other strains and was
373 characterized by the presence of a putative integrase gene located near a tRNA-Arg gene,
374 three genes encoding putative phage tail and lysis proteins, with the remaining genes having
375 unknown functions. Further, this region contained only co-oriented genes. Thus, we
376 hypothesize that this putative mobile genetic element (MGEx) encodes a prophage. In sum,
377 whereas isolate E27107 contained a shorter CRISPR array devoid of S22 and S23 and a
378 putative prophage harboring S22 and S23 protospacers, isolate E18064 harbored the two
379 spacers in its CRISPR array and lacked the corresponding prophage. Based on these
380 observations, we speculate that the presence of these two spacers provides resistance against

381 this particular MGE. The CRISPR and phage dynamics uncovered here emphasize the rapid
382 diversification of MGE-related genomic features in *Elizabethkingia* genomes, even at the
383 short evolutionary timescale of 5 years that separates the two African hospital infections.

384

385 CONCLUSIONS

386 *Elizabethkingia anophelis* was recently recognized as a cause of nosocomial
387 infections, and more recently, of community-acquired infections. In this work, the genome
388 sequences of two hospital meningitis African isolates were established and compared with
389 available genomes from other regions and sources. Our comparative phylogenetic analyses
390 led us to refine the identification of *Elizabethkingia* strains and to demonstrate firmly the
391 phylogenetic distinctness of the three described species of the genus. Comparative genomics
392 analysis revealed specific features of the African isolates including additional resistance
393 genes, a unique CRISPR locus and a specific putative capsular synthesis (*cps*) cluster.
394 Demonstration of the capsulated nature of the African *E. anophelis* isolates, and the presence
395 of variable *cps* clusters suggests an important determinant of the pathogenicity mechanisms
396 and virulence heterogeneity among *Elizabethkingia* strains. Analysis of the phylogenetic
397 structure of *E. anophelis* revealed several well-demarcated sublineages and demonstrated a
398 distinct evolutionary origin of African clinical isolates, excluding an epidemiological link
399 between recent African and Asian hospital infections. The core-genome MLST approach
400 defined here and made publicly available through the Institut Pasteur bacterial genome
401 database and analysis platform (<http://bigsdw.web.pasteur.fr/elizabethkingia/>) represents a
402 standardized high-resolution genotyping tool. It will enable to share a common language on
403 *E. anophelis* isolates sublineages, which will facilitate collaborative work on the population
404 biology, epidemiology and pathophysiology of this newly recognized bacterial species. Here,
405 the analysis of nosocomial infection isolates using this genotyping approach suggested that

406 the hospital environment, rather than mosquitoes, acts as a reservoir for *E. anophelis*, and that
407 the high persistence of the pathogen in this environment, if uncontrolled, allows for long-term
408 transmissions.

409

410 **MATERIALS AND METHODS**

411 **Isolates for genome sequencing**

412 The two clinical strains E18064 (alias: V0378064) and E27107 (alias: Po0527107)
413 isolated in Central African Republic were sequenced with a 100 base pair (bp) paired-end
414 protocol in Illumina HiSeq-2000. Libraries were constructed with the Nextera DNA Sample
415 Prep Kit (Illumina). Genome assembly was performed with the CLC Assembly Cell analysis
416 package version 3.2.2. The average number of contigs and the N50 statistic values (i.e., the
417 length for which half of the bases of a draft genome are situated in contigs of that length or
418 longer) observed from the two assembled genomes were similar to those of the publicly
419 available draft genomes (Table S1). The draft genome sequences of strains E18064 and
420 E27107 were annotated using the MicroScope/MaGe platform (41) and were deposited in the
421 European Nucleotide Archive under accession numbers CCAB000000000 and
422 CCAC000000000, respectively.

423

424 **Genomes from sequence database**

425 The eighteen *Elizabethkingia* genomic sequences available on April 1st, 2015 were
426 downloaded from the NCBI genome sequence repository (www.ncbi.nlm.nih.gov/genome).
427 The sequences comprised one complete genome and 17 whole-genome shotgun sequences
428 available as scaffolds or contigs. The strains were initially identified as *E. anophelis* (12
429 isolates), *E. meningoseptica* (n=4), *E. miricola* (n=1), and *Elizabethkingia* sp. (n=1) (Table
430 S1).

431

432 **Average nucleotide identity and 16S sequence analysis**

433 Average nucleotide identity values were estimated from the genomic assemblies using
434 JSpecies v1.2.1 with the BLASTN option (42). 16S rRNA sequences were gathered from
435 genomic sequences by BLASTN similarity search using the 16S sequence of the taxonomic
436 type strain R26^T (GenBank accession number EF426425) as a query.

437

438 **Core genome and pan-genome**

439 Core genomes (the set of genes present in all isolates) were built either for the genus or for
440 *E. anophelis* species. Orthologs were identified as bidirectional best BLASTP hits (43)
441 between the proteome of *E. anophelis* NUHP1 as a pivot and each of the other proteomes (18
442 for the genus and 15 for the species). Strain Endophthalmitis contained a large fraction of
443 pseudogenes (likely due to low quality of the sequence) and was therefore excluded from this
444 analysis. Hits with less than 40% (genus) or 80% (species) amino acid sequence similarity or
445 more than 20% difference in sequence length were discarded. Genomes from the same
446 species typically show low levels of genome rearrangements and this information can be used
447 to identify orthologs more accurately (44, 45). Therefore, the core-genome of the species was
448 defined as the intersection of pairwise lists of strict positional orthologs (as in (46)). The
449 core-genomes consist in the genes present in all genomes of the two sets. They were defined
450 as the intersection of the lists of orthologs between pairs of genomes.

451 Pan-genomes were built by clustering homologous protein-coding sequences (CDS)
452 into families. We determined the lists of putative homologs between pairs of genomes with
453 BLASTP used the E-values ($<10^{-4}$) to perform single-linkage clustering with SiLiX v1.2
454 (47). A CDS is thus included in a family if it shares a relation of homology to at least one
455 CDS already belonging to the family. SiLiX parameters were set to consider two CDS as

456 homologs if the aligned part showed at least 60% (*Elizabethkingia* genus) or 80%
457 (*E. anophelis*) identity and represented more than 80% of the smallest CDS length. The pan-
458 genomes thus represent the full complements of genes in the genus and in the species,
459 respectively. The pan-genomes of *Elizabethkingia* and of *E. anophelis* were determined
460 independently.

461

462 **Core genome multilocus sequence typing (MLST)**

463 To obtain a subset of CDSs that are highly reliable as genotypic markers, a subset of
464 the core genome was selected based on length conservation criterion leading to the selection
465 of 1,546 CDSs. These loci together constitute a cgMLST scheme useful for genotyping of
466 *E. meningoseptica*, *E. anophelis*, and *E. miricola* isolates. The scheme comprises fewer loci
467 than the core genome as expected from our more stringent selection criteria. The genome
468 sequences were then scanned for allelic variation using the BIGSdb tool (35) using the allele
469 sequences from reference strain R26 as the initial query sequences.

470

471 **Characterization of gene associated with drug resistance and horizontal gene transfer** 472 **mechanisms**

473 Acquired antimicrobial resistance (AR) genes were detected using HMMER3 (48)
474 against the ResFams, a curated database of AR protein families and associated profile hidden
475 Markov models, with the tblout and cut_ga options (49). Virulence-associated genes were
476 searched on the VFDB (50) using BLASTP (minimum 40% identity with E-value $<10^{-5}$), as
477 in (14).

478

479 **Comparison of genomes.** Local genomic alignments visualization was performed using
480 software BioNumerics v.7.5 (Applied-Maths, Sint-Martens Latem, Belgium) with default
481 seed matching and stretch extension parameters.

482

483 **Phylogenetic analyses**

484 For each of the 1,546 cgMLST loci, a multiple amino acid sequence alignment was
485 performed with MAFFT v.7.205 (default options) (51) that was back-translated in order to
486 obtain a codon-level alignment. The concatenation of these 1,546 multiple sequence
487 alignments was used to estimate the p-distance between each pair of isolates. These estimated
488 evolutionary distances were used to infer a genome-level phylogenetic tree with FastME
489 v.2.07 (Balanced Minimum Evolution criterion, SPR-based BME tree search) (52).

490 Phylogenetic tree analyses from single markers (i.e. 16S rRNA and CAT) were
491 performed from well-suited characters selected by BMGE v.1.12 (up to 50% allowed gap
492 proportion with models PAM1 and BLOSUM62, respectively) (53) from the multiple
493 sequence alignments generated by MAFFT. Tree inferences from 16S rRNA and CAT
494 sequences were performed by PhyML v.20131016 (54) (SPR-based ML tree search with
495 evolutionary models GTR+I and LG+ Γ 4+I, respectively).

496

497 **Detection of CRISPR-Cas systems**

498 Clusters of *cas* genes were identified and classified using MacSyFinder (27). To
499 identify *cas* pseudogenes, all Cas protein sequences previously detected were searched in all
500 the *Elizabethkingia* draft genomes using TBLASTN (E-value <10⁻³). CRISPR arrays were
501 identified following a previously published methodology (55). In short, they were identified
502 using CRT (CRISPR Recognition Tool) with the option -maxRL=60 (56), in all the
503 *Elizabethkingia* draft genomes. For each array, the repeats were extracted and were aligned

504 using MUSCLE (57). Then, we used *cons* (www.bioinformatics.nl/cgi-bin/emboss/help/cons)
505 to obtain consensus sequences from these multiple sequence alignments of each array. In all
506 cases, the consensus sequence corresponds to the most frequent sequence within a particular
507 array. We used the consensus sequence of the repeats as patterns to identify additional,
508 smaller and/or degenerate repeat clusters in all draft genomes with fuzznuc
509 (www.bioinformatics.nl/cgi-bin/emboss/help/fuzznuc). This step has also confirmed the lack
510 of any repeat in some draft genomes devoid of detectable CRISPR array and *cas* genes.

511

512 **Spacer diversity**

513 To identify additional spacers (i.e located at contig extremities) or to verify the lack of
514 a particular spacer in a given strain, all spacer sequences previously detected (i.e., flanked by
515 repeats at both sides) were searched in all the *Elizabethkingia* draft genomes using BLASTN.
516 Only hits showing at least 90% identity with the query, less than 20% difference in sequence
517 length and flanked by sequences similar to repeats were considered as additional spacers. We
518 then compared the spacer content within and between strains and species. Therefore, two
519 spacers were considered as similar (within strain) or common (between strains) if they had
520 less than 20% difference in sequence length and at least 90% identity.

521

522 **CRISPR locus orientation for polarized spacer acquisition**

523 During the adaptation phase, spacer acquisition occurs in a polarized fashion: new
524 spacers are typically integrated at the 5' (leader) end of the CRISPR array, which involves the
525 duplication of the first repeat of the array, and older spacers and repeats decay at the 3' end.
526 Consistent with this spacer acquisition model, we found that the most degenerated repeat was
527 located at one end of each CRISPR arrays. Hence, such particular repeat was considered as

528 the signature of the 3' end of the array, which corresponds to the oldest part of the locus.

529 Such positional information represents a timeline of spacer acquisition events.

530

531 **CRISPR locus PCR.** To confirm the genomic differences observed between CRISPR arrays

532 of the two African isolates based on Illumina sequence data, three primer pairs were designed

533 for PCR amplification and sequencing of the CRISPR locus in three overlapping fragments:

534 (i) L1F- CGT AAA CGT CTG TTA GAT GAT GG and sp15R-AAA CCA TTC TAC GGA

535 GAA C; (ii) sp17F-GAT GTA ATA AGA GTT GTT GCG and sp5R-TCG GAT TTA TGA

536 GGT GAT CCC; and (iii) sp7F-CAT AGA TCA CAC ATA CAG GGC and L1R-TGA GCG

537 CCC ATG TTG TCT CCG. PCR conditions for all amplification reactions were as follows:

538 initial denaturation at 94°C for 5 min; 30 cycles at 94°C for 30 s, 50°C for 30 s, and 72°C for

539 30s; and final extension at 72°C for 5 min. PCR products were purified by ultrafiltration

540 (Millipore), and nucleotide sequences were obtained using the PCR primers and BigDye

541 Terminator v1.1 chemistry (Applied Biosystems, Foster City, CA) on an ABI 3730XL

542 apparatus (Applied Biosystems, Foster City, CA). Sequence traces were edited and assembled

543 using BioNumerics.

544

545 **Detection of capsular gene clusters**

546 To identify capsular gene clusters, we performed a keyword search of the Pfam database

547 (pfam.xfam.org) for protein profiles involved in capsular polysaccharide production such as

548 glycosyl transferases, ABC transporters, *wzx* flippase and *wzy* polymerase (Table S6). We

549 then performed a search of these profiles in *Elizabethkingia* genomes using HMMER3 (48),

550 with the `tbl_out` option. This allowed us to clearly identify two capsule clusters in each of the

551 strains 502 and NBRC 12535. These clusters were both located in the same genomic region,

552 which included a highly conserved RecX family transcriptional regulator. We then searched

553 for *recX* in the other genomes and reconstructed the putative capsular polysaccharide
554 synthesis (*cps*) cluster in the *recX* neighborhood using the aforementioned protein profiles.

555 **FUNDING INFORMATION**

556 This study was supported by the Institut Pasteur (Programme Transversal de Recherche
557 CEVACAR No 385), by the CNRS and the European Research Council (Grant
558 EVOMOBILOME, number 281605), and by the French government's Investissement
559 d'Avenir program Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious
560 Diseases' (grant ANR-10-LABX-62-IBEID). O.R. is supported by an FRM fellowship (grant
561 number ARF20150934077).

562

563 **ACKNOWLEDGMENTS**

564 We would like to express our most sincere thanks to the director of Institut Pasteur in
565 Bangui, Dr Mirdad Kazanji, who made this work possible.

566

567 **REFERENCES**

- 568 1. **Bloch KC, Nadarajah R, Jacobs R.** 1997. *Chryseobacterium meningosepticum*: an
569 emerging pathogen among immunocompromised adults. Report of 6 cases and
570 literature review. *Medicine* **76**:30-41.
- 571 2. **Gungor S, Ozen M, Akinci A, Durmaz R.** 2003. A *Chryseobacterium*
572 *meningosepticum* outbreak in a neonatal ward. *Infect Control Hosp Epidemiol*
573 **24**:613-617.
- 574 3. **Adachi A, Mori T, Shimizu T, Yokoyama A, Takayama N, Ikeda Y, Okamoto S.**
575 2004. *Chryseobacterium meningosepticum* septicemia in a recipient of allogeneic cord
576 blood transplantation. *Scand J Infect Dis* **36**:539-540.
- 577 4. **Green O, Murray P, Gea-Banacloche JC.** 2008. Sepsis caused by *Elizabethkingia*
578 *miricola* successfully treated with tigecycline and levofloxacin. *Diagn Microbiol*
579 *Infect Dis* **62**:430-432.

- 580 5. **Kampfer P, Matthews H, Glaeser SP, Martin K, Lodders N, Faye I.** 2011.
581 *Elizabethkingia anophelis* sp. nov., isolated from the midgut of the mosquito
582 *Anopheles gambiae*. Int J Syst Evol Microbiol **61**:2670-2675.
- 583 6. **Frank T, Gody JC, Nguyen LB, Berthet N, Le Fleche-Mateos A, Bata P, Rafai C,**
584 **Kazanji M, Breurec S.** 2013. First case of *Elizabethkingia anophelis* meningitis in
585 the Central African Republic. Lancet **381**:1876.
- 586 7. **Teo J, Tan SY, Tay M, Ding Y, Kjelleberg S, Givskov M, Lin RT, Yang L.** 2013.
587 First case of *Elizabethkingia anophelis* outbreak in an intensive-care unit. Lancet
588 **382**:855-856.
- 589 8. **Lau SK, Wu AK, Teng JL, Tse H, Curreem SO, Tsui SK, Huang Y, Chen JH,**
590 **Lee RA, Yuen KY, Woo PC.** 2015. Evidence for *Elizabethkingia anophelis*
591 transmission from mother to infant, Hong Kong. Emerg Infect Dis **21**:232-241.
- 592 9. **Saez-Llorens X, McCracken GH, Jr.** 2003. Bacterial meningitis in children. Lancet
593 **361**:2139-2148.
- 594 10. **Gonzalez LJ, Vila AJ.** 2012. Carbapenem resistance in *Elizabethkingia*
595 *meningoseptica* is mediated by metallo-beta-lactamase BlaB. Antimicrob Agents
596 Chemother **56**:1686-1692.
- 597 11. **Kukutla P, Lindberg BG, Pei D, Rayl M, Yu W, Steritz M, Faye I, Xu J.** 2014.
598 Insights from the genome annotation of *Elizabethkingia anophelis* from the malaria
599 vector *Anopheles gambiae*. PLoS One **9**:e97715.
- 600 12. **Teo J, Tan SY, Liu Y, Tay M, Ding Y, Li Y, Kjelleberg S, Givskov M, Lin RT,**
601 **Yang L.** 2014. Comparative genomic analysis of malaria mosquito vector-associated
602 novel pathogen *Elizabethkingia anophelis*. Genome Biol Evol **6**:1158-1165.

- 603 13. **Bellais S, Aubert D, Naas T, Nordmann P.** 2000. Molecular and biochemical
604 heterogeneity of class B carbapenem-hydrolyzing beta-lactamases in
605 *Chryseobacterium meningosepticum*. *Antimicrob Agents Chemother* **44**:1878-1886.
- 606 14. **Li Y, Liu Y, Chew SC, Tay M, Salido MM, Teo J, Lauro FM, Givskov M, Yang**
607 **L.** 2015. Complete Genome Sequence and Transcriptomic Analysis of the Novel
608 Pathogen *Elizabethkingia anophelis* in Response to Oxidative Stress. *Genome Biol*
609 *Evol* **7**:1676-1685.
- 610 15. **Speer BS, Bedzyk L, Salyers AA.** 1991. Evidence that a novel tetracycline resistance
611 gene found on two *Bacteroides* transposons encodes an NADP-requiring
612 oxidoreductase. *J Bacteriol* **173**:176-183.
- 613 16. **Heir E, Lindstedt BA, Leegaard TM, Gjernes E, Kapperud G.** 2004. Prevalence
614 and characterization of integrons in blood culture *Enterobacteriaceae* and
615 gastrointestinal *Escherichia coli* in Norway and reporting of a novel class 1 integron-
616 located lincosamide resistance gene. *Ann Clin Microbiol Antimicrob* **3**:12.
- 617 17. **Chen YP, Tsao MY, Lee SH, Chou CH, Tsai HJ.** 2010. Prevalence and molecular
618 characterization of chloramphenicol resistance in *Riemerella anatipestifer* isolated
619 from ducks and geese in Taiwan. *Avian Pathol* **39**:333-338.
- 620 18. **Chen S, Bagdasarian M, Walker ED.** 2015. *Elizabethkingia anophelis*: molecular
621 manipulation and interactions with mosquito hosts. *Appl Environ Microbiol* **81**:2233-
622 2243.
- 623 19. **Moxon ER, Kroll JS.** 1990. The role of bacterial polysaccharide capsules as
624 virulence factors. *Curr Top Microbiol Immunol* **150**:65-85.
- 625 20. **Whitfield C, Paiment A.** 2003. Biosynthesis and assembly of Group 1 capsular
626 polysaccharides in *Escherichia coli* and related extracellular polysaccharides in other
627 bacteria. *Carbohydr Res* **338**:2491-2502.

- 628 21. **Li B, Zhao Y, Liu C, Chen Z, Zhou D.** 2014. Molecular pathogenesis of *Klebsiella*
629 *pneumoniae*. *Future Microbiol* **9**:1071-1081.
- 630 22. **Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen**
631 **HB, Nahm MH.** 2015. Pneumococcal Capsules and Their Types: Past, Present, and
632 Future. *Clin Microbiol Rev* **28**:871-899.
- 633 23. **Pan YJ, Lin TL, Chen CT, Chen YY, Hsieh PF, Hsu CR, Wu MC, Wang JT.**
634 2015. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79
635 capsular types of *Klebsiella* spp. *Sci Rep* **5**:15573.
- 636 24. **Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M,**
637 **McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM,**
638 **Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A,**
639 **Klugman KP, Parkhill J, Hanage WP, Bentley SD.** 2011. Rapid pneumococcal
640 evolution in response to clinical interventions. *Science* **331**:430-434.
- 641 25. **Ofek I, Kabha K, Athamna A, Frankel G, Wozniak DJ, Hasty DL, Ohman DE.**
642 1993. Genetic exchange of determinants for capsular polysaccharide biosynthesis
643 between *Klebsiella pneumoniae* strains expressing serotypes K2 and K21a. *Infect*
644 *Immun* **61**:4208-4216.
- 645 26. **Moxon ER.** 1992. Molecular basis of invasive *Haemophilus influenzae* type b
646 disease. *J Infect Dis* **165 Suppl 1**:S77-81.
- 647 27. **Abby SS, Neron B, Menager H, Touchon M, Rocha EP.** 2014. MacSyFinder: a
648 program to mine genomes for molecular systems with an application to CRISPR-Cas
649 systems. *PLoS One* **9**:e110726.
- 650 28. **Abby SS, Cury J, Guglielmini J, Neron B, Touchon M, Rocha EP.** 2016.
651 Identification of protein secretion systems in bacterial genomes. *Sci Rep* **6**:23080.

- 652 29. **McBride MJ, Nakane D.** 2015. *Flavobacterium* gliding motility and the type IX
653 secretion system. *Curr Opin Microbiol* **28**:72-77.
- 654 30. **Russell AB, Wexler AG, Harding BN, Whitney JC, Bohn AJ, Goo YA, Tran BQ,**
655 **Barry NA, Zheng H, Peterson SB, Chou S, Gonen T, Goodlett DR, Goodman**
656 **AL, Mougous JD.** 2014. A type VI secretion-related pathway in *Bacteroidetes*
657 mediates interbacterial antagonism. *Cell Host Microbe* **16**:227-236.
- 658 31. **Schwarz S, Hood RD, Mougous JD.** 2010. What is type VI secretion doing in all
659 those bugs? *Trends Microbiol* **18**:531-537.
- 660 32. **Jamet A, Nassif X.** 2015. New players in the toxin field: polymorphic toxin systems
661 in bacteria. *MBio* **6**:e00285-00215.
- 662 33. **Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F, Rocha EP.** 2011.
663 The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of
664 conjugation. *PLoS Genet* **7**:e1002222.
- 665 34. **Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard**
666 **AS, Garin B, Le Hello S, Arlet G, Nicolas-Chanoine MH, Decre D, Brisse S.**
667 2014. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella*
668 *pneumoniae* clonal groups. *Emerg Infect Dis* **20**:1812-1820.
- 669 35. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome
670 variation at the population level. *BMC Bioinformatics* **11**:595.
- 671 36. **Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA,**
672 **McCarthy ND.** 2013. MLST revisited: the gene-by-gene approach to bacterial
673 genomics. *Nat Rev Microbiol* **11**:728-736.
- 674 37. **Sorek R, Lawrence CM, Wiedenheft B.** 2013. CRISPR-mediated adaptive immune
675 systems in bacteria and archaea. *Annu Rev Biochem* **82**:237-266.

- 676 38. **Shariat N, Dudley EG.** 2014. CRISPRs: molecular signatures used for pathogen
677 subtyping. *Appl Environ Microbiol* **80**:430-439.
- 678 39. **Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de**
679 **Romans S, Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-**
680 **Jeanjean S, Achtman M, Brisse S, Sola C, Weill FX.** 2012. CRISPR typing and
681 subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One*
682 **7**:e36995.
- 683 40. **Chylinski K, Makarova KS, Charpentier E, Koonin EV.** 2014. Classification and
684 evolution of type II CRISPR-Cas systems. *Nucleic Acids Res* **42**:6091-6105.
- 685 41. **Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, Le Fevre F,**
686 **Longin C, Mornico D, Roche D, Rouy Z, Salvignol G, Scarpelli C, Thil Smith**
687 **AA, Weiman M, Medigue C.** 2013. MicroScope--an integrated microbial resource
688 for the curation and comparative analysis of genomic and metabolic data. *Nucleic*
689 *Acids Res* **41**:D636-647.
- 690 42. **Richter M, Rossello-Mora R.** 2009. Shifting the genomic gold standard for the
691 prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**:19126-19131.
- 692 43. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman**
693 **DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database
694 search programs. *Nucleic Acids Res* **25**:3389-3402.
- 695 44. **Dandekar T, Snel B, Huynen M, Bork P.** 1998. Conservation of gene order: a
696 fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**:324-328.
- 697 45. **Rocha EP.** 2006. Inference and analysis of the relative stability of bacterial
698 chromosomes. *Mol Biol Evol* **23**:513-522.
- 699 46. **Touchon M, Hoede C, Tenailon O, Barbe V, Baeriswyl S, Bidet P, Bingen E,**
700 **Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O,**

- 701 **Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L,**
702 **Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S,**
703 **Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z,**
704 **Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Medigue C, Rocha EP,**
705 **Denamur E.** 2009. Organised genome dynamics in the *Escherichia coli* species
706 results in highly diverse adaptive paths. *PLoS Genet* **5**:e1000344.
- 707 47. **Miele V, Penel S, Duret L.** 2011. Ultra-fast sequence clustering from similarity
708 networks with SiLiX. *BMC Bioinformatics* **12**:116.
- 709 48. **Eddy SR.** 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**:e1002195.
- 710 49. **Gibson MK, Forsberg KJ, Dantas G.** 2015. Improved annotation of antibiotic
711 resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*
712 **9**:207-216.
- 713 50. **Chen L, Xiong Z, Sun L, Yang J, Jin Q.** 2012. VFDB 2012 update: toward the
714 genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids*
715 *Res* **40**:D641-645.
- 716 51. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software
717 version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772-780.
- 718 52. **Desper R, Gascuel O.** 2004. Theoretical foundation of the balanced minimum
719 evolution method of phylogenetic inference and its relationship to weighted least-
720 squares tree fitting. *Mol Biol Evol* **21**:587-598.
- 721 53. **Criscuolo A, Gribaldo S.** 2010. BMGE (Block Mapping and Gathering with
722 Entropy): a new software for selection of phylogenetic informative regions from
723 multiple sequence alignments. *BMC Evol Biol* **10**:210.
- 724 54. **Guindon S, Gascuel O.** 2002. Efficient biased estimation of evolutionary distances
725 when substitution rates vary across sites. *Mol Biol Evol* **19**:534-543.

- 726 55. **Touchon M, Rocha EP.** 2010. The small, slow and specialized CRISPR and anti-
727 CRISPR of *Escherichia* and *Salmonella*. PLoS One **5**:e11126.
- 728 56. **Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P.**
729 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered
730 regularly interspaced palindromic repeats. BMC Bioinformatics **8**:209.
- 731 57. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced
732 time and space complexity. BMC Bioinformatics **5**:113.
- 733
- 734
- 735

736 **FIGURES AND TABLES**

737 **FIG 1.** Heat-map of the ANI_b for each pairwise comparison (a) and intra and inter-species
738 ANI_b variability (b). Green dashed line on the top dendrogram marks groups of taxa that are
739 from the same species or have more than 95% ANI_b values and therefore might be regarded
740 as part of the same species. Black asterisks indicate strains previously described as
741 *E. meningoseptica*. Black rectangles indicate the two hospital-acquired isolates from Central
742 African Republic sequenced in this study.

743

744 **FIG 2.** Core- and pan-genome sizes of *Elizabethkingia anophelis* (a) and spectrum of
745 frequencies for *E. anophelis* gene repertoires (b). The pan- and core-genomes were used to
746 perform gene accumulation curves. These curves describe the number of new genes (pan-
747 genome) and genes in common (core-genome) obtained by adding a new genome to a
748 previous set. The procedure was repeated 1,000 times by randomly modifying the order of
749 integration of genomes in the analysis. The spectrum of frequencies represents the number of
750 genomes where the families of the pan-genome can be found, from 1 for strain-specific genes
751 to 16 for core genes. Blue indicates accessory genes and green the genes that are highly
752 persistent in *E. anophelis*. Numbers in circles above the bars indicate the number of
753 antimicrobial resistance (AR) gene families.

754

755 **FIG 3.** Phylogenetic tree of the *Elizabethkingia* genus based on the multiple sequence
756 alignments of the 1,542 cgMLST loci. Black dots correspond to bootstrap supports higher
757 than 90%.

758

759 **FIG 4.** Genetic organization of the putative capsular polysaccharide synthesis (*cps*) gene
760 clusters in *Elizabethkingia* genomes. Open reading frames (ORFs) are shown as rectangles

761 and are drawn to scale. Intergenic spaces are not to scale. Three highly conserved genes,
762 including *recX*, are depicted in brown. Green indicates highly conserved genes involved in
763 capsule translocation (*capD*, *wza*, *wzc*). Yellow and red indicate genes *wzx* and *wzy*,
764 respectively.

765

766 **FIG 5.** CRISPR-Cas type II-C organization and diversity. (a) System organization is
767 depicted on the top, with *cas* genes in brown, and the repeat-spacer array in black. Below the
768 gene scheme, the repeat and spacer (captured MGEs) content is detailed as black diamonds
769 and white rectangles, respectively. L corresponds to the leader sequence. Bottom line,
770 consensus repeat sequence and degenerated repeat sequence in the 5 genomes containing
771 such system. (b) Diversity of spacer content between the two Central African Republic
772 hospital isolates. (c) Presence of a large MGE containing perfect matches with the two
773 additional spacers S22 and S23.

774

775 **Supplementary tables**

776 **TABLE S1.** General characteristics of the 20 *Elizabethkingia* genomes analyzed in this
777 study.

778

779 **TABLE S2.** Average nucleotide identity (58) values computed from 20 *Elizabethkingia*
780 genome sequences.

781

782 **TABLE S3.** Antimicrobial resistance-associated features identified in *Elizabethkingia*
783 genomes.

784

785 **TABLE S4.** Predicted virulence factors identified in *Elizabethkingia* genomes.

786

787 **TABLE S5.** Protein secretion systems identified in *Elizabethkingia* genomes.

788

789 **TABLE S6.** Proteins families searched in *Elizabethkingia* genomes to characterize capsular
790 polysaccharide synthesis clusters.

791

792 **Supplementary figures**

793 **FIG S1.** Phylogenetic analysis of *Elizabethkingia* isolates based on 16S rRNA sequences. a.

794 Phylogenetic tree with bootstrap-based support (500 replicates) >70% indicated at branches.

795 b. Detail of the multiple 16S rRNA sequence alignment (aligned characters 963-1,033).

796

797 **FIG S2.** Gene tree of chloramphenicol acetyltransferase (CAT) genes. Blue: sequences from

798 Schwarz et al. 2004, FEMS Microbiology Reviews 28:519–542; magenta: sequences from

799 *Elizabethkingia* strains; black: closely related homologous sequences selected from RefSeq.

800 NCBI accession ids are given on the right. Bootstrap-based supports (500 replicates) >70%

801 are indicated at branches.

802

803 **FIG S3.** Distribution of CRISPR-Cas systems in *Elizabethkingia* genus. White indicates

804 absence of the trait and black indicates its presence. The numbers reported on the right

805 correspond to the number of spacers identified in each genome.

806

807 **FIG S4.** Gene-by-gene comparison of the *Elizabethkingia* genomes based on 1,542 loci

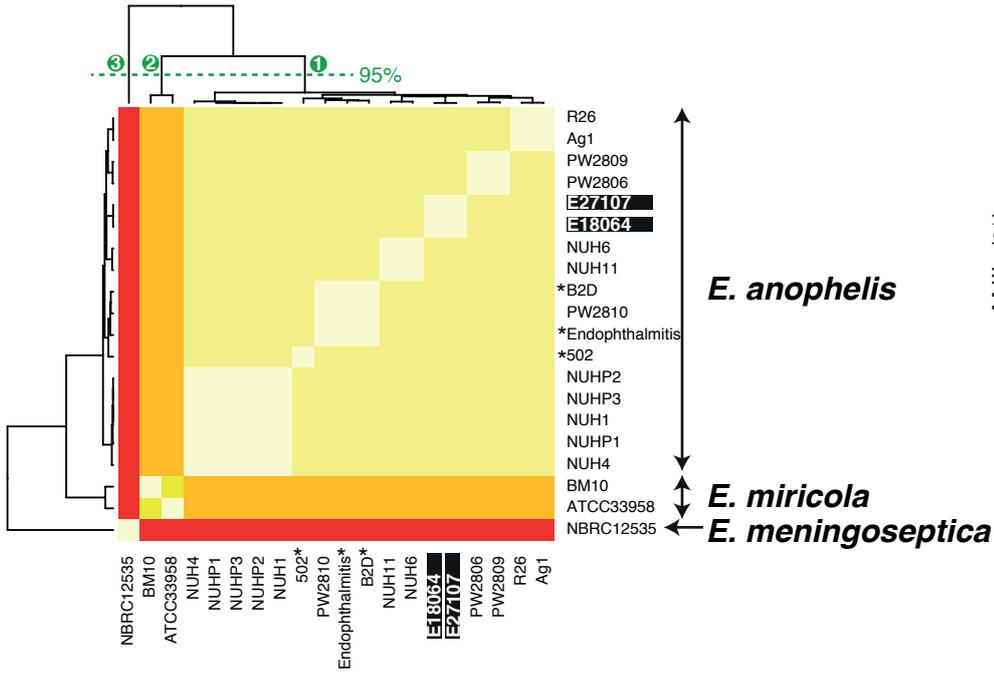
808 included in the core genome multilocus sequence typing (cgMLST) scheme. a: UPGMA

809 dendrogram computed from the matrix of allelic mismatches distances (i.e., the proportion of

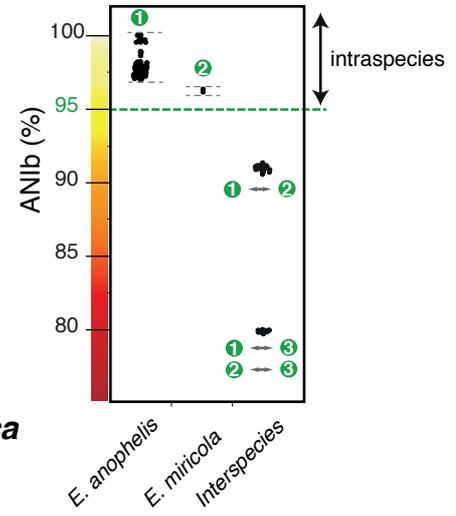
810 distinct alleles among each pair of profiles). b: Minimum spanning tree of the Singapore

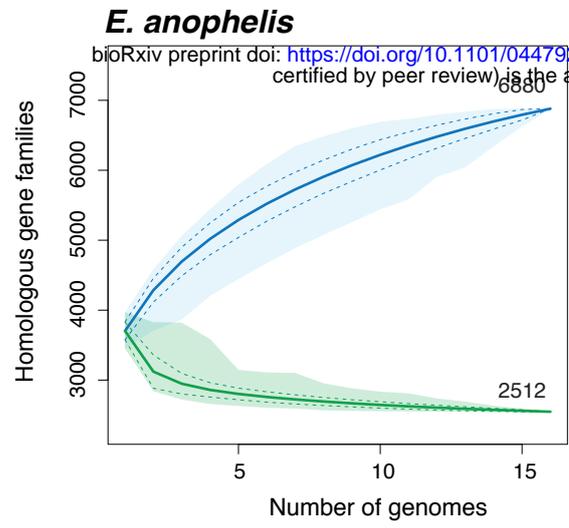
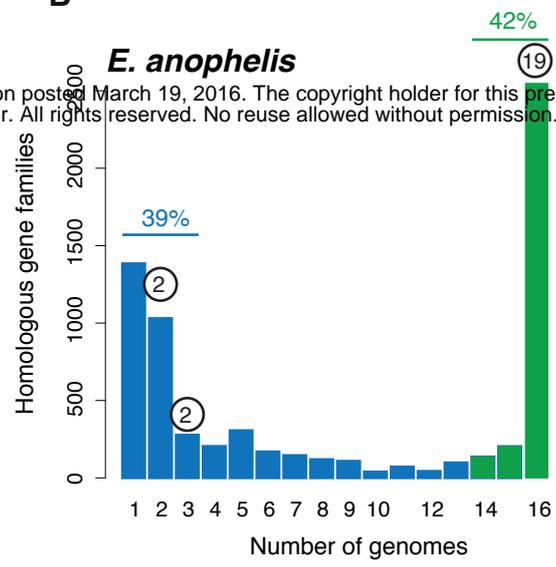
- 811 lineage. The number of allelic mismatches between each displayed link is given. c: Minimum
812 spanning tree of the African lineage, which differ by 4 alleles out of 1,542.

a

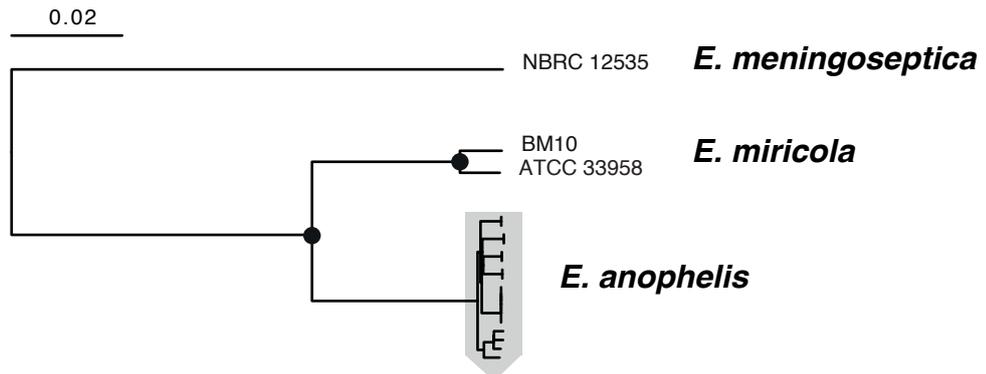


b

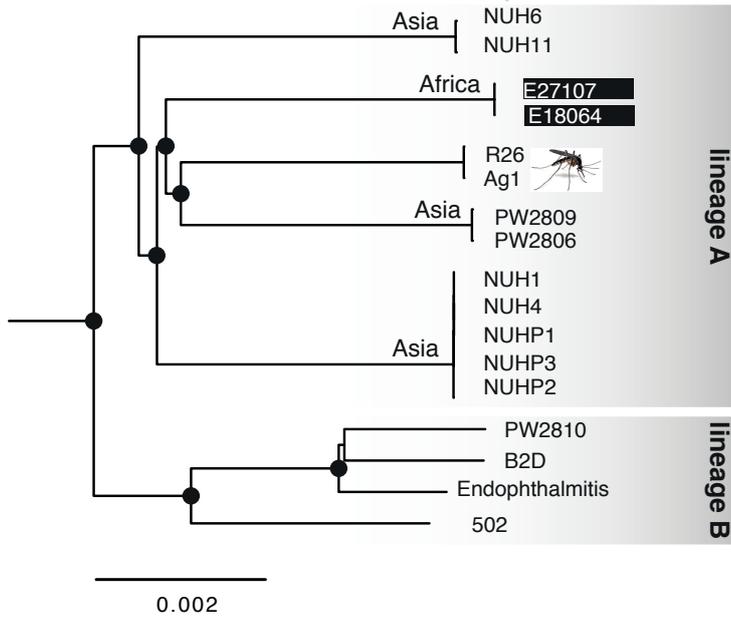


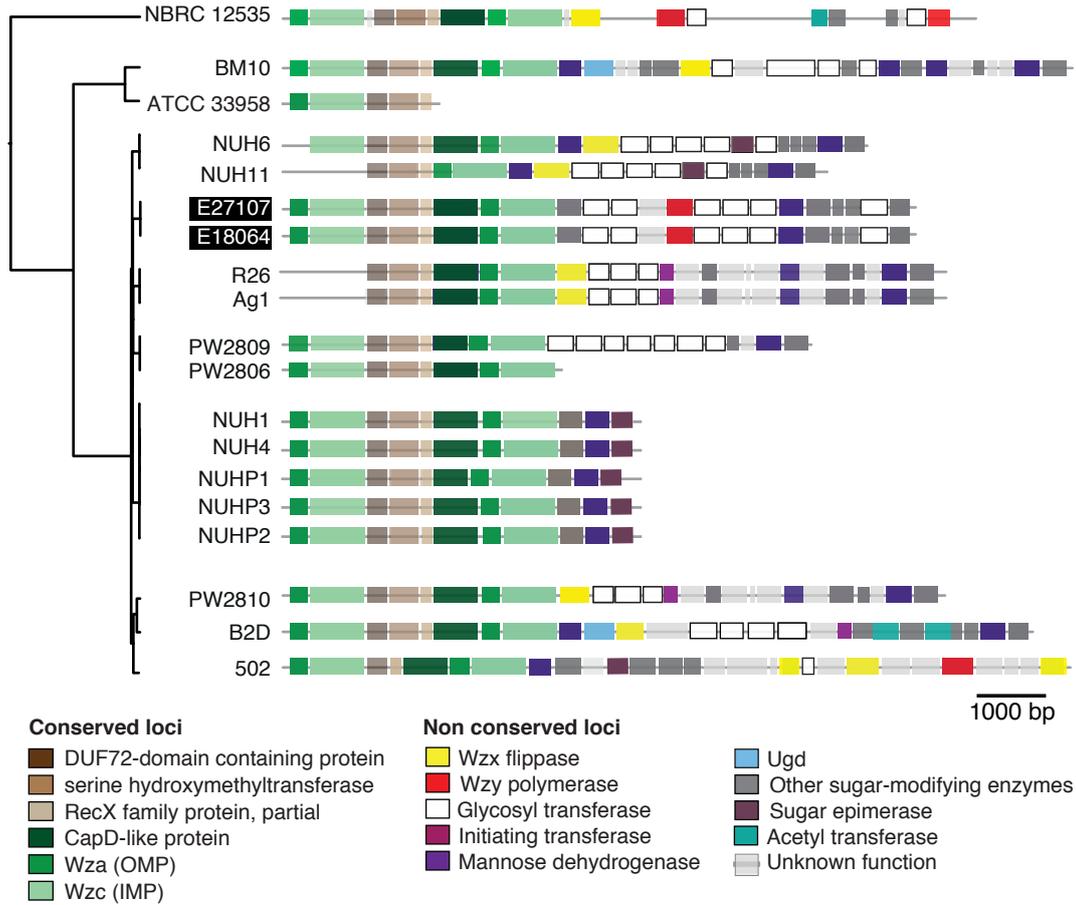
A**B**

a



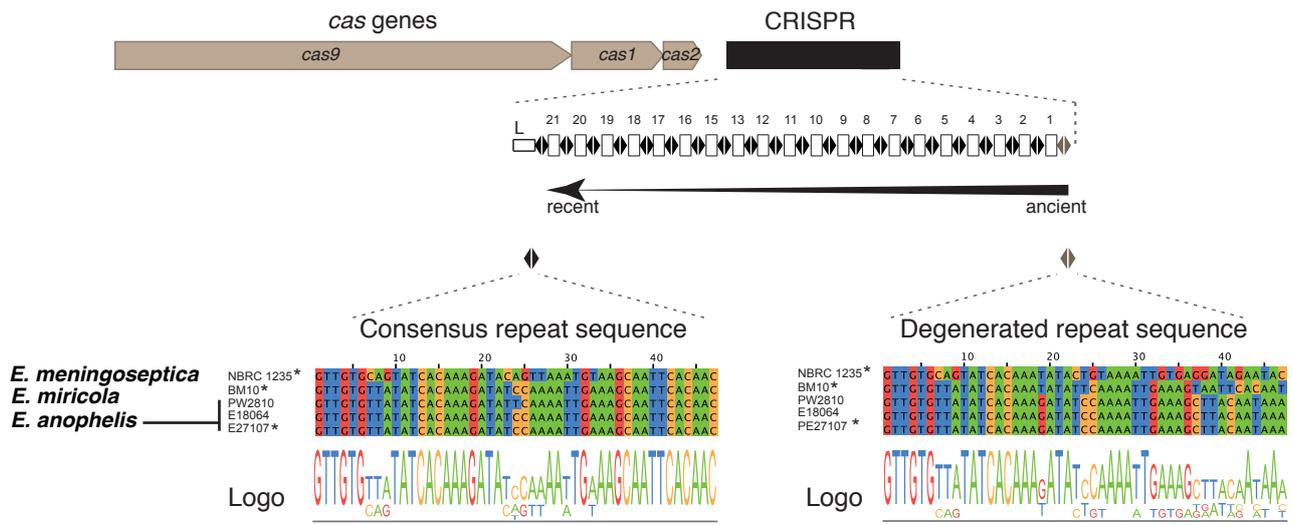
b



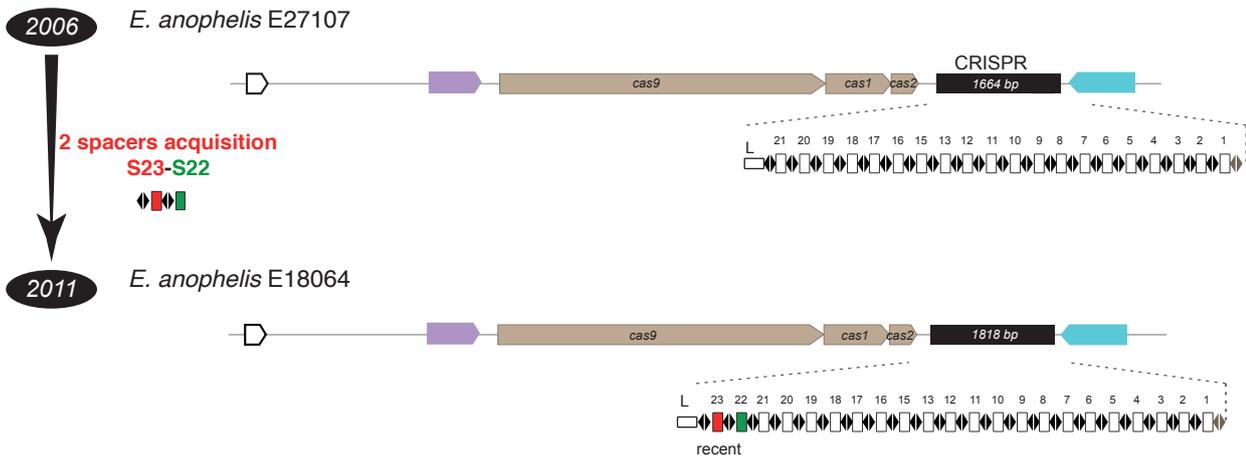


a

CRISPR-Cas Type II-C



b



c

