A global *Corynebacterium diphtheriae* genomic framework sheds light on current diphtheria reemergence

Authors

Melanie Hennart [a,b,c], Chiara Crestani [a], Sebastien Bridel [a], Nathalie Armatys [a,b], Sylvie Brémont [a,b], Annick Carmi-Leroy [a,b], Annie Landier [a,b], Virginie Passet [a,b], Laure Fonteneau [e], Sophie Vaux [e], Julie Toubiana [a,b,d], Edgar Badell [a,b] and Sylvain Brisse [a,b,*]

Affiliations

[a] Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, F-75015, Paris, France

[b] Institut Pasteur, National Reference Center for Corynebacteria of the Diphtheriae Complex, Paris, France

[c] Sorbonne Université, Collège doctoral, F-75005 Paris, France

[d] Department of General Pediatrics and Pediatric Infectious Diseases, Hôpital Necker-Enfants Malades, APHP, Université de Paris, Paris, France

[e] Santé publique France, Saint-Maurice, France

***Correspondence**:

Sylvain Brisse: Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, 25-28 rue du Docteur Roux, F-75724, Paris, France; Phone: +33 1 45 68 83 34 ; E-mail: sylvain.brisse@pasteur.fr

Keywords: diphtheria, genomic sequencing, antimicrobial resistance, virulence, epidemiology, transmission, 2022 reemergence, bioinformatics tool

Running Title: Genomic surveillance of diphtheria using DIPHTOSCAN

Abstract

Background

Diphtheria, caused by *Corynebacterium diphtheriae*, reemerges in Europe since 2022. Genomic sequencing can inform on transmission routes and genotypes of concern, but currently, no standard approach exists to detect clinically important genomic features and to interpret emergence in the global *C. diphtheriae* population framework.

Methods

We developed the bioinformatics pipeline DIPHTOSCAN (available at https://gitlab.pasteur.fr/BEBP/diphtoscan) to extract from genomes of Corynebacteria of the *diphtheriae* species complex, medically relevant features including *tox* gene presence and disruption. We analyzed 101 human *C. diphtheriae* isolates collected in 2022 in metropolitan and overseas France (France-2022). To define the population background of this emergence, we sequenced 379 additional isolates (mainly from France, 2018-2021) and collated 870 publicly-available genomes.

Results

The France-2022 isolates comprised 45 *tox*-positive (44 toxigenic) isolates, mostly imported, belonging to 10 sublineages (<500 distinct core genes). The global dataset comprised 245 sublineages and 33.9% *tox*-positive genomes, with DIPHTOSCAN predicting non-toxigenicity in 16.0% of these. 12% of the global isolates, and 43.6% of France-2022 ones, were multidrug resistant. Convergence of toxigenicity with penicillin and erythromycin resistance was observed in 2 isolates from France-2022. Phylogenetic lineages Gravis and Mitis contrasted strikingly in their pathogenicity-associated genes.

Conclusions

This work provides a bioinformatics tool and global population framework to analyze *C. diphtheriae* genomes, revealing important heterogeneities in virulence and resistance features. Emerging genotypes combining toxigenicity and first-line antimicrobial resistance represent novel threats. Genomic epidemiology studies of *C. diphtheriae* should be intensified globally to improve understanding of reemergence and spatial spread.

Introduction

Diphtheria was a leading cause of infant mortality before the implementation of anti-toxin therapy and mass vaccination programs. Classical diphtheria is a respiratory infection mainly caused by the *tox* gene-positive strains of the bacterium *Corynebacterium diphtheriae.* The disease is classically characterized by the presence of pseudomembranes on the tonsils, pharynx and larynx. Only some strains of *C. diphtheriae* can produce the diphtheria toxin, which is encoded by the *tox* gene carried by a prophage integrated into the chromosome of these strains. The toxigenic strains can induce severe systemic symptoms that include myocarditis and peripheral neuropathies. Other forms of infection include bacteriemic infections, most often caused by non-toxigenic strains, and cutaneous infections, which are considered to play an important role in the transmission of the pathogen.

Diphtheria has been virtually eliminated by mass vaccination, but can cause large outbreaks where vaccination coverage is insufficient (du Plessis et al., 2017; Polonsky et al., 2021; Badell et al., 2021). In France, no case was reported between 1990 and 2001 (Bonmarin et al., 2009), and in the 2017-2021 period only 6.4 *tox*-positive *C. diphtheriae* were detected per year by the French surveillance (our unpublished data). In striking contrast, in 2022, 45 *tox*-positive isolates were detected, including 34 from metropolitan France, mostly associated with recent arrival from abroad. *C. diphtheriae* also reemerges in several European countries, strongly associated with non-vaccinated young adults with cutaneous infections with a travel history from Afghanistan and other countries (Badenschier et al., 2022; Kofler et al., 2022).

Whole genome sequencing (WGS) is a powerful approach to understand transmission and define the pathogenicity-associated characteristics of infectious isolates. *C. diphtheriae* is a genetically diverse species with multiple phylogenetic sublineages among which a large heterogeneity of virulence or antimicrobial resistance factors is observed (Sangal & Hoskisson, 2016; Seth-Smith & Egli, 2019; Hennart et al., 2020; Guglielmini et al., 2021). One prominent polymorphism in *C. diphtheriae* is the variable presence of the *tox* gene, but the population dynamics and drivers of *tox* acquisition or loss remain poorly understood. In addition, non-toxigenic *tox*-bearing (NTTB) *C. diphtheriae* isolates represent 5-20% of *tox*-positive isolates, but our capacity to predict toxigenicity from genomic sequences is still limited. Several other experimentally-demonstrated virulence factors have been described in *C. diphtheriae* (Ott, 2018). Although early 1930s literature suggested a higher virulence of isolates of biovar Gravis (McLeod, 1943; Barksdale, 1970), it is unknown whether this historical observation applies to extant diphtheria cases, as recent Gravis isolates are more rarely *tox*-positive than those of biovar Mitis (Hennart et al., 2020). More generally, the population variation of virulence factors, and its interactions with clinical outcomes, remain largely to be characterized. Despite being rare, antimicrobial resistance (AMR) in *C. diphtheriae* is increasingly reported (Mina et al., 2011; Zasada, 2014; Forde et al., 2020; Hennart et al., 2020), but the mechanisms of resistance that are prevalent across world regions are not well known, and the evolutionary emergence and dissemination of multi-drug resistant *C. diphtheriae*, and its possible convergence with toxigenicity in the same strains, should be carefully monitored.

Although WGS of *C. diphtheriae* clinical isolates is increasingly performed for surveillance purposes, no simple tool currently exists for *C. diphtheriae* genomic feature extraction and interpretation in clinical,

surveillance and research contexts. Besides, analyses of *C. diphtheriae* genomes remain largely unstandardized, which limits the interpretation of local genomic epidemiology studies in their global context. Advances towards standardization include the 7-gene MLST genotyping approach and attached nomenclature of sequence types (ST) (Bolt et al., 2010), and its core-genome MLST (cgMLST) extension and associated nomenclature of sublineages and genomic clusters (Guglielmini et al., 2021).

Here, we aimed to provide insights into the France 2022 diphtheria emergence by reporting on its epidemiology and by placing the involved isolates in the global genomic context of *C. diphtheriae* populations. We introduce DIPHTOSCAN, a genotyping tool designed for rapid and standardized genomic analyses of Corynebacteria of the *C. diphtheriae* species complex (CdSC), and illustrate its use by analyzing the 101 *C. diphtheriae* isolates (including 56 *tox*-negative ones) collected in 2022 in France (henceforth, the France-2022 dataset). We provide context of this emergence by analyzing 1249 other *C. diphtheriae* genomes of diverse geographic and temporal origins, including 379 newly sequenced isolates collected by the French national surveillance laboratory, mostly between 2018 and 2021. We uncovered novel insights into the global population structure of *C. diphtheriae*, including a striking contrast in pathogenesis-associated gene clusters between phylogenetic lineages Gravis and Mitis, and describe high-risk sublineages with convergence of resistance and virulence features.

Material & Methods

Clinical isolates inclusion and global genomic sequence dataset

To investigate the epidemiology of diphtheria in France, we included all cases of *C. diphtheriae* infections detected by the French surveillance in 2022. Among 144 isolates received by the National Reference Center, there were 101 deduplicated isolates when retaining only one from each patient. These were isolated in metropolitan France as well as in Mayotte, La Reunion and French Guiana (**France-2022 dataset, Table S1**). Note that metropolitan France comprises mainland France and Corsica, as well as nearby islands in the Atlantic Ocean, the English Channel (French: la Manche), and the Mediterranean Sea. All isolates collected in 2022 from metropolitan France were from mainland France. Overseas France is the collective name for all the French territories outside Europe.

In addition, a total of 1,249 comparative genomes were included (**Table S1**). First, we sequenced for the present study 379 additional isolates, including 320 collected prospectively between 2008 and 2021 by the French National Reference Center (NRC), 34 historical clinical isolates mostly from metropolitan France and 19 isolates from Algeria (Benamrouche et al., 2016). These new genomes were sequenced to complement the 226 previous genomes from *C. diphtheriae* from the French diphtheria surveillance system (Hennart et al., 2020; Guglielmini et al., 2021), including 43 isolates from Yemen (Badell et al., 2021). Together, these represent 599 produced by the NRC for Corynebacteria of the *diphtheriae* complex (**non-2022 French NRC dataset, Table S1**). Nearly four-fifths (531; 88.6%) of these isolates were prospectively collected between 2008 and 2021 from French metropolitan and overseas territories, 54 isolates (9.0%) were collected between 1990 and 2007 from France and Algeria and 14 (2.3%) isolates collected between 1951 and 1987 from metropolitan France.

132 Second, we included publicly-available genomes from NCBI, mostly previously published and
133 isolated in South Africa (du Plessis et al., 2017), Germany-Switzerland (Meinel et al., 2016), Germany
134 (Dangel et al., 2018; Berger et al., 2019), Canada (Chorlton et al., 2019), Austria (Schaeffer et al., 2020), the
135 USA (Williams et al., 2020; Xiaoli et al., 2020), Spain (Hoefer et al., 2020), India (Will et al., 2021) and
136 Australia (Timms et al., 2018). Altogether, this represents a dataset of 579 genomes (**non-French public**
137 **dataset, Table S1**).
138 Further, we sequenced 6 ribotype reference strains (Grimont et al., 2004). Together with 65
139 previously sequenced (Hennart et al., 2020), this represents a dataset of 71 genomes of ribotype reference
140 strains (**Table S1**).
141 From the global set of 1,249 genomes (**non-2022 French NRC + non-French public dataset +**
142 **ribotype datasets**), we created a non-redundant subset of genomes by randomly selecting one genome per
143 genomic cluster (threshold: 25 cgMLST mismatches; see below), isolation year and city (if city was
144 unavailable, the country was used instead); this deduplicated subset comprised 976 genomes (hereafter,
145 the ***global dataset***).
146
147 **Microbiological characterization of isolates at the French National Reference Laboratory**
148 *C. diphtheriae* isolates were grown and purified on Tinsdale agar. Strains were characterized
149 biochemically for pyrazinamidase, urease, and nitrate reductase and for utilization of maltose and trehalose
150 using API Coryne strips (BioMérieux, Marcy l'Etoile, France) and the Rosco Diagnostica reagents (Eurobio,
151 Les Ulis, France). The Hiss serum water test was used for glycogen fermentation. The biovar of isolates was
152 determined based on the combination of nitrate reductase (positive in Mitis and Gravis, negative in Belfanti)
153 and glycogen fermentation (positive in Gravis only). Antimicrobial susceptibility was determined by disc
154 diffusion (BioRad, Marnes-la-Coquette, France). Zone diameter interpretation breakpoints are given in
155 **Table S3**.
156 The presence of the diphtheria toxin *tox* gene was determined by real-time PCR assay (Badell et
157 al., 2019), whereas the production of the toxin was assessed using the modified Elek test (Engler et al.,
158 1997).
159 For genomic sequencing, isolates were retrieved from −80°C storage and plated on tryptose-casein
160 soy agar for 24 to 48 h. A small amount of bacterial colony biomass was resuspended in a lysis solution
161 (20 mM Tris-HCl [pH 8], 2 mM EDTA, 1.2% Triton X-100, and lysozyme [20 mg/ml]) and incubated at 37°C
162 for 1 h. DNA was extracted with the DNeasy Blood&Tissue kit (Qiagen, Courtaboeuf, France) according to
163 the manufacturer's instructions. Genomic sequencing was performed using a NextSeq500 instrument
164 (Illumina, San Diego, CA) with a 2 × 150-nucleotide (nt) paired-end protocol following Nextera XT library
165 preparation (Hennart et al., 2020).
166 For de novo assembly, paired-end reads were clipped and trimmed using AlienTrimmer v0.4.0 (Criscuolo &
167 Brisse, 2013), corrected using Musket v1.1 (Liu et al., 2013), and merged (if needed) using FLASH
168 v1.2.11(Magoč & Salzberg, 2011). For each sample, the remaining processed reads were assembled and
169 scaffolded using SPAdes v3.12.0 (Bankevich et al., 2012).
170

171 **Merging of the Oxford and Pasteur MLST databases**

172 Two *C. diphtheriae* databases using the BIGSdb framework were originally designed separately for
173 distinct purposes: while Oxford's PubMLST database mainly offered 7-gene MLST (Bolt et al., 2010), the
174 Pasteur database was used for the *Corynebacterium* cgMLST typing (Guglielmini et al., 2021). To facilitate
175 the use of these resources and avoid redundancy in the curation of the two independent genomic libraries,
176 a merging of the databases was decided in agreement with PubMLST administrators. In order to merge the
177 data available in the two databases, we proceeded as per BIGSdb dual design: isolates genomes and
178 provenance data were imported into the "isolates" database, whereas allelic definitions of MLST were
179 imported into the "seqdef" database.

180 Regarding the isolates database, we first downloaded Oxford's PubMLST *C. diphtheriae* database.
181 To avoid isolate entries duplication, we identified common isolates between the two databases, and filtered
182 duplicate isolates before import into the Pasteur database. In total, 684 out of 934 (73%) isolates from the
183 Oxford database were imported. To facilitate the tracing of isolates and their possible previous existence in
184 Oxford's database, isolates identification numbers (BIGSdb-Pasteur ID number) of isolates from the Oxford
185 database were numbered from 1,520 to 2,003. We also collated them into a public project collection called
186 "Oxford" (project ID 13).

187 Regarding the sequence and profiles definition database, we imported MLST alleles and profiles
188 into an initially void MLST scheme container within the BIGSdb-Pasteur database. MLST analysis was
189 performed on all isolates of the BIGSdb-Pasteur database, including the ones imported from Oxford, which
190 were therefore assigned the same MLST genotype as previously in the Oxford database.

191 At the end of the merging process, all isolates and MLST data from PubMLST's *C. diphtheriae*
192 database were available into the BIGSdb-Pasteur *C. diphtheriae* species complex database
193 (https://bigsdb.pasteur.fr/diphtheria/), and Oxford's PubMLST *C. diphtheriae* database was shut down. As
194 of September 22nd, 2022, the database resulting from the merged datasets comprised 1,478 public isolates
195 records with 794 associated genomes, and 2,392 isolates in total when considering private entries. The
196 number of entries varied across species: *C. diphtheriae* (n = 1,291; 87.4%) and *C. ulcerans* (n = 131; 8.9%),
197 *C. belfantii* (n = 45; 3.0%) and *C. rouxii* (n = 10; 0.7%). The MLST scheme comprised 854 registered STs.

198

199 **cgMLST and nomenclature of sublineages**

200 The MLST and cgMLST genotypes (cgST) were defined using the Institut Pasteur *C. diphtheriae*
201 species complex database at https://bigsdb.pasteur.fr/diphtheria.

202 A core genome MLST (cgMLST) scheme comprising 1,305 loci (Guglielmini et al., 2021) was
203 employed to define the alleles and cgST of the 1,249 genomic sequences using BIGSdb
204 (https://bigsdb.pasteur.fr/diphtheria). Using the 1,249-genomes dataset, the mean number of missing
205 alleles per profile was 12 (0.9%) and almost all (n=1,242; 99.4%) genomes had a cgMLST profile with fewer
206 than 65 (5%) missing alleles. A cgST number was defined for all but one cgMLST profiles (one genome had
207 219 missing alleles, whereas the admissible threshold is 10%, i.e., 130 missing alleles).

208 Genomes were classified using the single-linkage cluster-profile.pl function of BIGSdb into genomic
209 clusters (25 mismatch threshold) and sublineages (500 mismatches). Sublineages were attributed numbers

210 by using an ST inheritance rule (Hennart et al., 2022), which was applied from SL1 to SL744, after which the
211 numbers are attributed consecutively with no reference to MLST identifiers, starting at 10,000 (see column
212 'SL' in **Table S1**).
213
214 **Phylogenetic analysis based on a core genome**
215 Panaroo v1.2.3 was used to generate from the assembled genomic sequences, a core genome used
216 to construct a multiple sequence alignment (cg-MSA). The genome sequences were first annotated using
217 prokka v1.14.5 with default parameters, resulting in GFF files. Protein-coding gene clusters were defined
218 with a threshold of 70% amino acid identity, and core genes were concatenated into a cg-MSA when present
219 in 95% of genomes. IQtree version 2 was used to build a phylogenetic tree based on the cg-MSA, with the
220 best fitting model TVM+F+R5. The tree was constructed from 1,948 core genome loci, for a total alignment
221 length of 1,986,172 bp (79.8% of NCTC13129 genome length, of 2,488,635 bp), was rooted using *C. belfantii*
222 strain FRC0043^{T}, and is available at: https://itol.embl.de/tree/1579917435471751662784292.
223
224 **Development of the DIPHTOSCAN pipeline**
225 To develop DIPHTOSCAN, we combined code from Kleborate v2.2.0 (Lam et al., 2021), AMRfinderPlus
226 (Feldgarden et al., 2021) and BIGSdb (Jolley & Maiden, 2010) with some modifications. The structures of
227 DIPHTOSCAN and its custom database are presented in **Figure S3** and **Figure S4**. A custom code was created
228 for DIPHTOSCAN initiation, interpretation and for displaying results. The *C. diphtheriae* specific genes
229 (genomic markers, AMR determinants and virulence factors) for which the genomes are screened by
230 DIPHTOSCAN (**Figure S4**) are provided in a custom database similar in its structure to the AMRFinderPlus
231 database (https://ftp.ncbi.nlm.nih.gov/pathogen/Antimicrobial_resistance/); this database can be further
232 enriched with novel features in the future. When launching DIPHTOSCAN, the AMRFinderPlus and custom
233 databases are merged. We used the functions of species determination, MLST genotyping, and full CDS
234 prediction from Kleborate. All the functionalities are presented in **Figure S2**. To facilitate readability and
235 downstream analyses, the output of DIPHTOSCAN is generated in a tab-delimited format. The execution time
236 of DIPHTOSCAN increases linearly with the number of input genomes. Roughly, 40 seconds are needed to
237 scan a single genome with 1 cpu. DIPHTOSCAN computations can be parallelized, as AMRFinderPlus and
238 JolyTree use parallelization.
239
240 **Assignment of species, MLST and Sequence Types (ST)**
241 To perform rapid and accurate species identification, DIPHTOSCAN uses the k-mer-derived Mash
242 distances (Ondov et al., 2016). DIPHTOSCAN calculates Mash distances (Mash v2.2) between the query
243 genomes and a collection of reference assemblies of the *CdSC*, and reports the species with the smallest
244 distance. *C. diphtheriae* genomes were confirmed as *C. diphtheriae* based on a Mash distance smaller than
245 0.05 with either the *C. diphtheriae* type strain NCTC11397^{T} (= C7S), the reference genome strain
246 NCTC13129, or the vaccine strain PW8 (Park-Williams 8).
247 Mash distance ≤0.05 is reported as a strong match, ≤0.1 as weak. We have used and adapted the
248 structure of the Kleborate tool for this function. This approach was validated by comparing DIPHTOSCAN

249 species assignments with those obtained by average nucleotide identity (ANI; Konstantinidis and Tiedje,
250 2005) using FastANI (Jain et al., 2018) using the global dataset; 100% concordance was achieved.
251 MLST profiles and sequence types (ST) were defined using the international MLST scheme for *C.*
252 *diphtheriae* and *C. ulcerans*. DIPHTOSCAN defines these genotypes for genomic sequences using the
253 analogous script from Kleborate. In order to use an up-to-date version of the MLST nomenclature, which is
254 regularly updated, the MLST profiles and alleles are downloaded at the start of the pipeline before
255 genotyping the genomes. The download_alleles.py script from BIGSdb is used for this purpose
256 (https://github.com/kjolley/BIGSdb/tree/develop/scripts/rest_examples).
257
258 **Biovar-associated markers detection**
259 The three main biovars of *C. diphtheriae* can be distinguished based on isolate abilities to reduce nitrate
260 and to metabolize glycogen. Previously, a strong concordance was found between the biovar and the
261 presence in the genome of several genomic markers including *spuA*, which codes for a putative alpha-1,6-
262 glycosidase, and the *narKGHJI* operon for nitrate reductase (Sangal et al., 2014; Santos et al., 2018; Hennart
263 et al., 2020). We therefore included in the custom DIPHTOSCAN query database the *spuA* marker and its
264 adjacent genes (DIP0351; DIP0353; DIP0354; DIP0357=*spuA*), which are strongly associated with biovar
265 Gravis, and the *narIJHGK* cluster, which is typically absent or partly disrupted, mainly due to mutations in
266 the *narG* (Hennart et al., 2020) or *narI* (Sangal et al., 2014) in isolates of biovar Belfanti. In the future,
267 markers of the two biovars of *C. pseudotuberculosis* may be added.
268
269 **Detection of antibiotic resistance genes**
270 Antibiotic resistant genes were identified using AMRfinderPlus, with the database found at:
271 https://ftp.ncbi.nlm.nih.gov/pathogen/Antimicrobial_resistance/. Features are detected by using the
272 BLAST family of tools, with identity and coverage defined for each family of antibiotics (fam.tab). A few
273 genes particularly relevant for the *CdSC* were added to this database: *pbp2m* (Forde et al., 2020; Hennart
274 et al., 2020) and mutation points of *rpoB* (WP_004566675.1) and *gyrA* (WP_010933942.1). AMRfinderPlus
275 v3.11.2 is used within DIPHTOSCAN with no modifications.
276
277 **Detection of virulence genes from the *C. diphtheriae* species complex**
278 A custom database of virulence features of *C. diphtheriae* and related species was compiled from
279 literature for the purposes of this work. We included in the custom query database, a panel of genetic
280 features for which published experimental evidence of their clinical relevance exists in *C. diphtheriae* or
281 closely related species (*i.e.,* increased virulence in animal models, or decreased antimicrobial susceptibility
282 *in vitro*) (**Table S2**). These target genes are the following: *tox*, SpaA-, SpaD-, and SpaH-type pili gene clusters,
283 DIP0733 (*67-72p*), the genes DIP1281 and DIP1621 that code for proteins of the NlpC/P60 family, DIP0543
284 (*nanH*), DIP1546 and DIP2093 (Ott, 2018) and *pld* (phospholipase). A second panel of genetic features with
285 no experimental evidence but with strong suspicion for a role in virulence, based on homology with genes
286 from other pathogens, was also included for broader screening of virulence features (**Table S2**).

287 For the main virulence factor, the *tox* gene, we used a reference sequence of this gene from each
288 of *C. diphtheriae*, *C. ulcerans* and *C. pseudotuberculosis* (WP_003850266.1, WP_014835773.1 and
289 WP_014654963.1, respectively), as the toxin differs between these species (Dangel et al., 2019).
290 The *tox* gene may be disrupted in some strains by the occurrence of stop codons or other genetic
291 events, leading to non-toxigenic, *tox*-gene bearing (NTTB) isolates (Zakikhany et al., 2014; Melnikov et al.,
292 2022). DIPHTOSCAN provides information on the putative toxicity of a strain from the *tox* gene sequence using
293 a categorization into four possible outputs, following the convention proposed in Kleborate (Lam et al.,
294 2021): (i) if the sequence in the analyzed genome is identical to the reference tox sequence from
295 NCTC13129 strain, the output provides the name of the sequence with the denomination of the species
296 (*e.g.,* tox_diphtheriae); (ii) If the sequence in the analyzed genome has a coverage length identical to the
297 reference, but an identity different from 100%, then an asterisk (*) is added (*e.g.,* tox_diphtheriae*); (iii) If
298 the hit coverage length is smaller than the reference length, the tag '-NTTB?- xx%' is added, where xx is the
299 percentage of the missing sequence length compared to the reference length); (iv) Finally, if the truncated
300 *tox* sequence is located at the end of a contig, the symbol '$' is added, to highlight that the prediction is
301 uncertain.
302 To analyse the *tox* gene promotor region, the sequence of strain NCTC13129 corresponding to the
303 300 nt upstream of *tox* start codon was used as a query in BLASTn analyses. Sequence alignment of the
304 corresponding region in the queried genomes was performed with seaView and the mutations were
305 visualized and compared with the distribution of the available results of the Elek test. For DtxR sequence
306 variation, *dtxR* was detected using BLASTn with DIPHTOSCAN, and the translation into amino acid, alignment
307 and visualization of mutations were performed using seaView.
308 Virulence genes were identified using the method of AMRfinderPlus but based on our custom
309 database of virulence features. The virulence genes are detected by BLASTn with thresholds of minimum
310 80% identity and 50% coverage. Based on the output of AMRfinderPlus, the gene completion and allele
311 similarity is reported as described above for the *tox* gene following the Kleborate convention.
312

313 **Results**

314 **1. The re-emergence of *C. diphtheriae* in France in 2022**

315 In 2022, the French NRC has received 101 human samples of *C. diphtheriae,* from metropolitan France
316 (n=76) as well as in the Indian Ocean islands of Mayotte (n=10) and La Reunion (n=6), and in French Guiana
317 (n=9). There were 45 isolates carrying the *tox* gene coding for diphtheria toxin (*tox*-positive isolates),
318 whereas in the five previous years a total of 32 *tox*-positive *C. diphtheriae* were detected (**Figure S1A**).
319 *C. diphtheriae* were isolated in metropolitan France (n=34) and in Mayotte/La Reunion (n=11), while none
320 were found in French Guiana. The metropolitan France isolates were isolated only in the second part of the
321 year (**Figure S1B**) and were associated with a recent travel history from Afghanistan (n=24) or other
322 countries from West Africa, North Africa, Middle East and Southern Asia; These isolates were
323 predominantly from cutaneous infections, whereas 7 were from respiratory infections (**Table S1**; **Figure 1**).
324 Only 3 of the 34 patients were up to date with their vaccination.

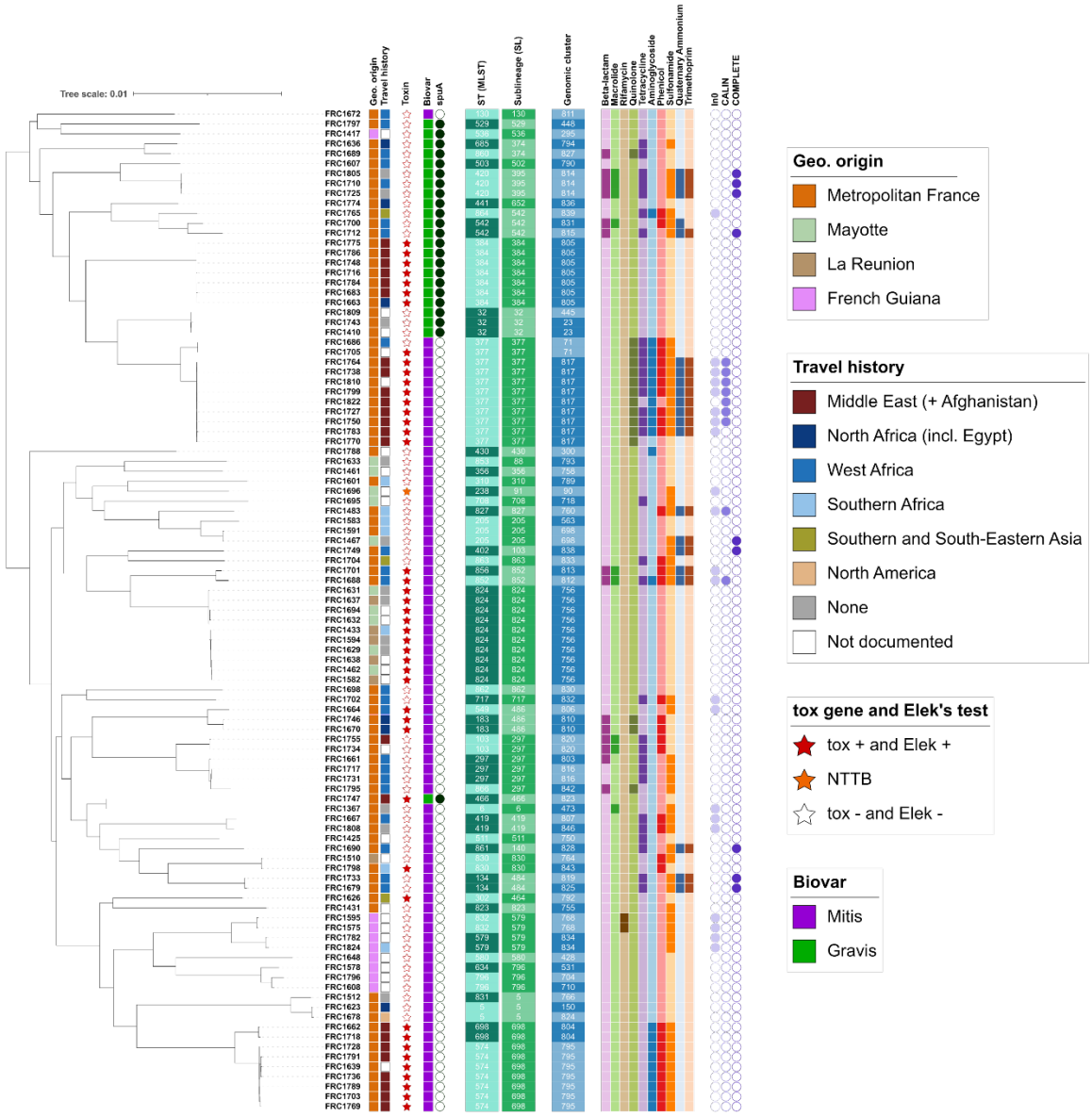

Figure 1. Phylogenetic tree of *Corynebacterium diphtheriae* from France, 2022

The tree was obtained by maximum likelihood based on a multiple sequence alignment of the core genome. The scale bar represents the number of nucleotide substitutions per site. The first column that follows the isolates identifiers indicates the geographic origin (place of isolation; see key). Travel history provides the most distant geographic region of reported travel (see key); note that Afghanistan was included in Near and Middle East; and Egypt was included in North Africa. The stars represent the presence (red star), presence but disruption (NTTB, orange) or absence (white star) of the diphtheria toxin tox gene. Biovars are represent in colored squares, and spuA gene presence by a dark green circle. MLST STs, sublineage (SL) and genomic clusters are provided with an alternation of colored strips. Identifiers of the main STs are indicated (note the strong concordance between ST and cgMLST sublineages). The 10 next colored columns correspond to the presence of at least one gene or mutation (for quinolone and rifamycin classes) involved in resistance to the indicated class of antimicrobial agents. Last, the presence of integron-related structures *(Cury et al., 2016)* is indicated: In0 (integron integrase and no attC sites), CALIN (clusters of attC sites lacking integron-integrases) and complete integrons (integrase and at least one attC site). The simultaneous presence of In0 and CALIN may denote their presence in different contigs even though the integron might be complete.

341 **2. Development of the DIPHTOSCAN pipeline**

342 To provide a tool to extract information from genomes of *C. diphtheriae* and related potentially
343 toxigenic species, we developed DIPHTOSCAN. The technical characteristics of DIPHTOSCAN are summarized in
344 **Figure S2-S4** and the methodological details for genotyping are provided in the Methods section.

345 In brief, the DIPHTOSCAN pipeline (**Figure S2**) starts with taxonomic assignment of species. Recent
346 taxonomic updates have defined, besides the three classical species *C. diphtheriae*, *C. ulcerans* and
347 *C. pseudotuberculosis*, three novel species of the Corynebacteria of the *diphtheriae* species complex (CdSC):
348 *C. belfantii* (Dazas et al., 2018), *C. rouxii* (Badell et al., 2020) and *C. silvaticum* (Dangel et al., 2020). If the
349 genome is confirmed to belong to the CdSC, 7-gene MLST analysis (Bolt et al., 2010) is performed. For
350 *C. diphtheriae*, additional genotype categorizations can be performed using the BIGSdb-Pasteur database
351 tool: cgST, genomic cluster and sublineage assignment (Guglielmini et al., 2021). Next, the detection of
352 antimicrobial resistance determinants (mutations in core genes and horizontally acquired genes) and
353 virulence factors is performed. DIPHTOSCAN also includes a prediction of the functionality or disruption of
354 the *tox* gene, the most important virulence factor of CdSC isolates. DIPHTOSCAN next searches for genomic
355 markers associated with biovars Gravis, Mitis and Belfanti, a biochemical-based classification that was
356 initiated in the 1930s (Anderson et al., 1931; McLeod, 1943) and which is still in use for *C. diphtheriae* strain
357 characterization. IntegronFinder2 (Néron et al., 2022) was included in the pipeline to contextualize
358 resistance genes. Last, a rapid phylogenetic method based on k-mer distances, JolyTree (Criscuolo, 2020),
359 was integrated to provide quick phylogenetic trees for the genomic assembly datasets under study. The
360 two latter steps are optional.

361

362 **3. Genetic diversity of *C. diphtheriae* isolates from France, 2022**

363 The *C. diphtheriae* isolates belonging to the France-2022 dataset were sequenced and their genomic
364 sequences were analyzed using DIPHTOSCAN. Sublineage classification of the isolates showed that the France-
365 2022 dataset comprised 41 distinct sublineages (defined using the 500 cgMLST mismatch threshold). The
366 nomenclature of these sublineages was established using an inheritance rule that captures their majority
367 MLST denomination, where possible (Guglielmini et al., 2021; Hennart et al., 2022), resulting in a strong
368 concordance of sublineage denominations with the classical MLST identifiers (**Figure 1**). There were 51
369 different STs, as 9 sublineages comprised two or more closely related STs; in 7 of 9 cases, they only differed
370 by a single locus. Sublineages thus appeared as useful classifiers for closely related STs.

371 There were four frequently isolated *tox*-positive sublineages: SL824 included 10 isolates from Mayotte
372 and La Reunion; these all belonged to the same genomic cluster (GC756), indicating recent transmission.
373 Three other frequent *tox*-positive sublineages were SL377 (n=11 isolates, 10 of which were *tox*-positive),
374 SL698 (n=9) and SL384 (n=7), which were associated with travel from Afghanistan and countries of the
375 Middle East (**Figure 1**). Whereas SL384 was genetically homogeneous (GC805), SL377 and SL698 both
376 comprised two genomic clusters (SL377: GC817 and GC71; SL698: GC795-ST574 and GC804-ST698). SL377-
377 GC71 was not associated with Afghanistan and one isolate from Senegal was *tox*-negative.

378 Besides the above four frequent sublineages, six additional *tox*-positive sublineages were isolated:
379 three isolates of sublineage SL486 associated with Senegal and Tunisia; two SL852 isolates associated with

380 Mali; and one SL466 isolate associated with travel from Afghanistan and one SL464 isolate associated with
381 Thailand. SL91 comprised one non-toxigenic, *tox*-gene bearing (NTTB) isolate, and SL830 comprised 2
382 isolates: one *tox*-positive and one *tox*-negative.
383 Besides, there were 31 *tox*-negative sublineages, which were typically isolated once or twice only; a
384 notable exception was SL297, which comprised six *tox*-negative isolates associated with travel from Egypt,
385 Senegal, and Mali (**Figure 1**).
386
387 **4. The global phylogenetic framework of *C. diphtheriae***
388 We investigated the global diversity of *C. diphtheriae* to provide context to the France-2022
389 emerging genotypes. A dataset of 1,249 comparative *C. diphtheriae* genomes were sequenced or gathered
390 from previous studies (see Methods). cgMLST grouped these isolates into 245 sublineages. The 7-gene
391 MLST analysis revealed 364 distinct STs. Almost all (360; 98.6%) STs corresponded one-to-one with the
392 sublineage level, *i.e.*, all isolates of these STs belonged to the same sublineage. However, 72 sublineages
393 (29.4%) comprised at least two STs. Of the 123 novel sublineages uncovered here, 114 sublineages were
394 given an identifier inherited from the 7-gene MLST nomenclature (whereas 9 were attributed an arbitrary
395 number, see Methods).
396 There were 576 genomic clusters, many of which comprised previously documented epidemiological
397 clusters of related isolates. For example, GC456 comprised 43 isolates from a Vancouver inner city outbreak
398 (Chorlton et al., 2019). Whereas 47 GCs had between 5 and 27 isolates (**Table S1**; **Figure S5A**), the 529
399 remaining ones had only 1 and 4 isolates. 106 (43.3%) of the 245 sublineages comprised at least two
400 genomic clusters.
401 To eliminate the population bias introduced by multiple sampling of outbreak strains, we created a
402 non-redundant subset by randomly selecting one genome per genomic cluster, isolation year and city (if
403 city was unavailable, the country was used instead) and with the same resistance genes profile and *tox*
404 status (see column 'Dataset' in **Table S1**). These 976 deduplicated genomes (hereafter, the *global dataset*)
405 define the background population of *C. diphtheriae*.
406 Within the global dataset, 35 sublineages were represented 7 times of more (**Figure 2**). The two
407 predominant sublineages were SL8 (n=61) and SL5 (n=48); their main 7-gene MLST sequence types were
408 ST8 and ST5, previously noted to be predominant in the ex-USSR 1990s outbreak. The most represented
409 *tox*-positive sublineages in the global dataset were SL8, SL453, SL486, SL377 and SL91, and SL50 was a
410 predominant NTTB sublineage (**Figure 2**).
411 Of the 10 sublineages with *tox*-positive isolates observed in France-2022, 7 were found in the global
412 dataset; of which 5 were among the 35 frequent global sublineages. Besides, 9 *tox*-negative sublineages
413 from France-2022 were also frequent in the global dataset (**Figure 2**). Of the common France-2022
414 sublineages, SL377, SL384 and SL297 were also common in the global dataset (**Figure 2**), and their
415 toxigenicity and resistance features matched those observed in the global dataset. In contrast, SL698
416 (metropolitan France) and SL824 (Indian Ocean) were uniquely common in the France-2022 dataset (**Figure**
417 **S5B**).

418 The phylogenetic structure of *C. diphtheriae* revealed a star-like phylogeny with multiple deeply-
419 branching sublineages as previously reported (Berger et al., 2019; Seth-Smith & Egli, 2019; Hennart et al.,
420 2020; Guglielmini et al., 2021) (**Figure 3**). Sublineages appeared to be grouped according to biovars Gravis
421 (and its *spuA* marker gene) and Mitis as previously noted (Hennart et al., 2020), as they formed two main
422 lineages named Gravis (green branches) and Mitis (purple), defined by the presence of the *spuA* gene (**Table**
423 **S1**). cgMLST-defined sublineages were highly concordant with the phylogeny and often comprised more
424 than one 7-gene ST (**Figure 3; Table S1**). The frequent *tox*-positive sublineages SL377 and SL384 were
425 phylogenetically related within lineage Gravis (**Figure 3**), suggesting they share ancestrally-acquired genetic
426 features.
427 We placed within this population background, the France-2022 isolates (**Figure S6**), which appeared to
428 be dispersed in multiple branches of the global phylogeny. The isolates previously collected by the French
429 reference laboratory appeared even more diverse and largely dispersed across the global phylogenetic
430 diversity of *C. diphtheriae* (**Figure S6**), indicating that a large fraction of the global diversity has been
431 sampled by the French surveillance system.
432 Ribotyping was previously used as a classification and nomenclature system of *C. diphtheriae* strains
433 (Grimont et al., 2004; Mokrousov, 2009). The 71 ribotype reference strains sequenced herein or previously
434 (Hennart et al., 2020) were placed in the global phylogeny (**Figure S7**), showing that these strains are highly
435 diverse. However, this ribotype subset is biased towards tox-positives (40 of 71 strains) and appears to
436 represent unevenly and incompletely, the currently sampled *C. diphtheriae* diversity.

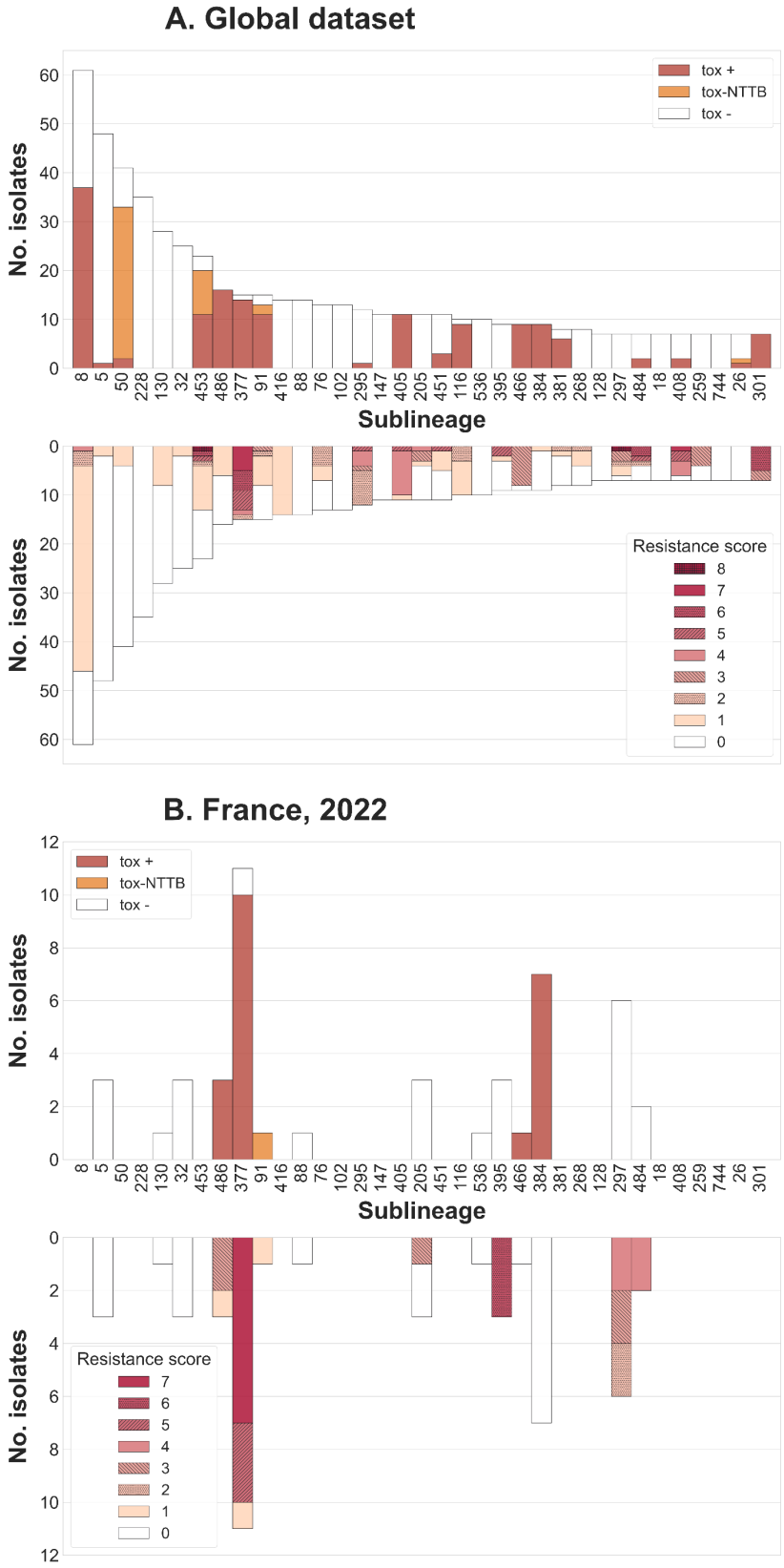

Figure 2. Sublineage distribution of tox gene and resistance score

(Top) Bar length correspond to the number of isolates per sublineage (deduplicated global dataset, 976 isolates). Upper part: isolates with non-disrupted tox are colored in red, with disrupted tox (NTTB) in orange, and not carrying the tox gene in white. Lower part: bar sectors are colored by resistance score (including beta-lactams and macrolides; see key).

(Bottom) Bar length correspond to the number of isolates per sublineage (France, 2022 dataset, 101 isolates). Bar sectors are colored as in the top panel.

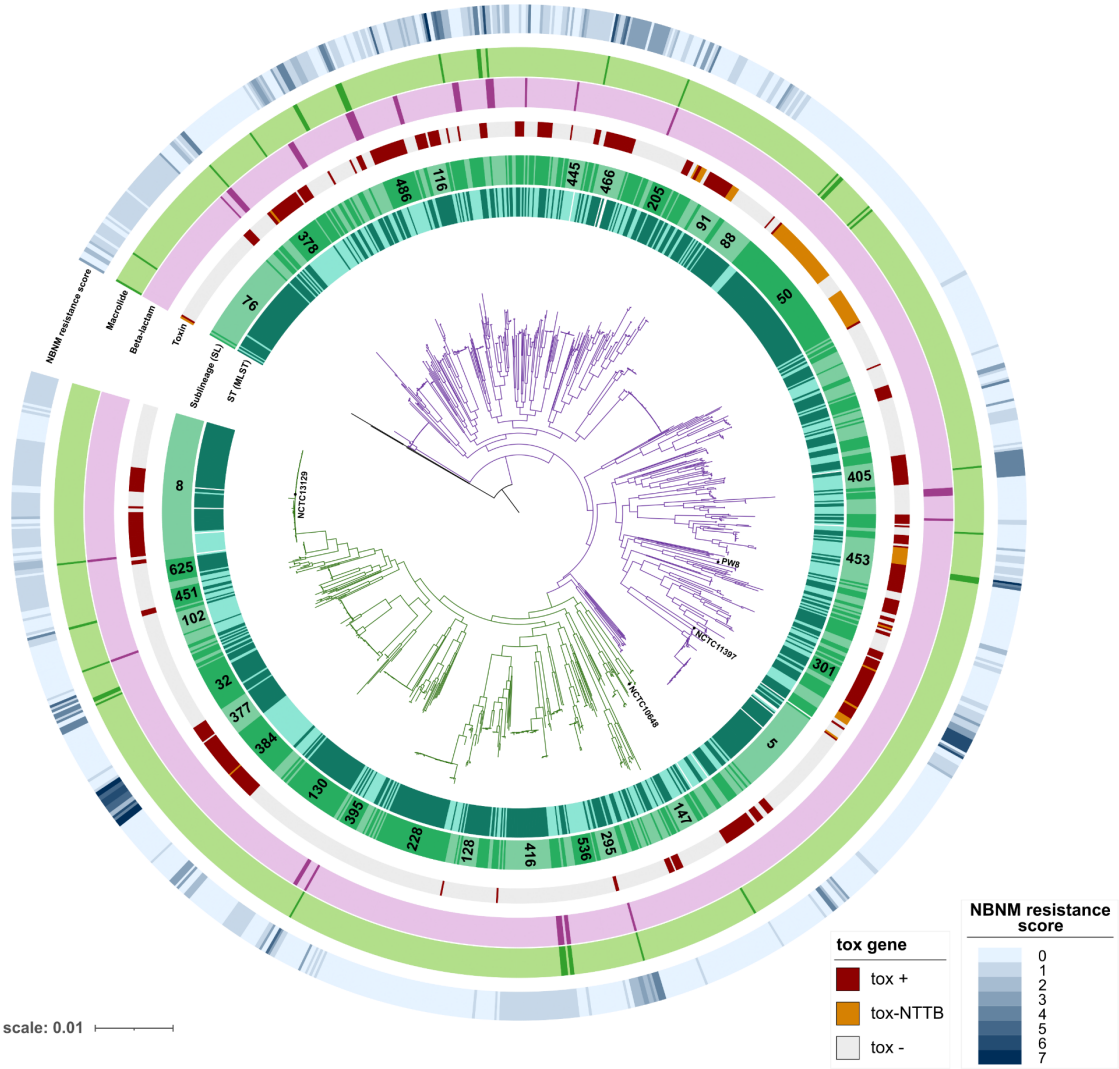

Figure 3. Phylogenetic tree of *Corynebacterium diphtheriae*

The tree was obtained by maximum likelihood based on a multiple sequence alignment of the core genome, and was rooted with C. belfantii (not shown). The scale bar gives the number of nucleotide substitutions per site. The main lineages Mitis and Gravis are drawn using purple and green branches, respectively. The two inner circles indicate MLST and sublineage alternation, respectively; main sublineages are labeled within the sectors. first ten colored circles around the tree correspond to the different classes of antibiotics. The following circle indicates the presence, disruption or absence of the diphtheria toxin tox gene (see key). The beta-lactam resistance circle indicates the presence of the pbp2m gene, while the macrolide circle corresponds to the presence of *ermX* or *ermC* (darker color: presence of the genomic determinant). The most external circle indicates the non-beta-lactam, non-macrolide (NBNM) resistance score (number of classes with at least one resistance feature), as a blue gradient (see key). Four reference strains are indicated: strain NCTC13129, which is used as genomic sequence reference; strain NCTC10648, which is used as the tox-positive and toxinogenic reference strain in PCR and Elek tests, respectively; strain NCTC11397^{T}, which is the taxonomic type strain of the C. diphtheriae species; and the vaccine production strain PW8.

5. Population distribution of the diphtheria toxin gene

To evaluate DIPHTOSCAN for its ability to detect the *tox* gene and to predict its toxigenicity, we used the 855 isolates for which data on *tox* qPCR and Elek test were available. DIPHTOSCAN detected that *tox* was located at the end of a contig and therefore incomplete in 3 cases (reported with a '$' suffix, indicating genomic assembly truncation). Of the 852 remaining isolates, 221 were *tox*-positive and 631 *tox*-negative

464 by the reference qPCR method. DIPHTOSCAN detected the *tox* gene in 219 (99.1%) of the *tox*-positives, and
465 reported its absence in 2 isolates. Among the 631 *tox*-negative isolates, DIPHTOSCAN reported the absence
466 of the gene in 625 (99.0) isolates. Of 198 Elek-positives, 195 (98.5%) were predicted to be toxigenic by
467 DIPHTOSCAN, whereas 1 was predicted to be non-toxigenic and for two isolates the *tox* gene was not
468 detected. Of the Elek-negative isolates, 11 (50.0%) were predicted as non-toxigenic by DIPHTOSCAN. Thus,
469 *tox* detection by DIPHTOSCAN was both sensitive and specific, whereas toxigenicity prediction was highly
470 sensitive but not highly specific, likely due to unexplained non-toxigenicity in isolates with a full-length toxin
471 gene. However, we did not observe non-toxigenicity-associated variation in the promoter region of the *tox*
472 gene, nor on the DtxR protein sequence.
473 In the France 2022 dataset, 45 genomes were detected as *tox*-positive and 44 of these were predicted
474 as toxigenic, with 100% concordance with the Elek test. In comparison, within the global dataset,
475 approximately one third of the isolates (331/976; 33.9%) were *tox*-positive, as defined using DIPHTOSCAN,
476 which detected a truncation and hence predicted non-toxigenicity in 16.0% of these (52/331).
477 Combining the global dataset with the France 2022 dataset (1077 genomes), DIPHTOSCAN identified 33
478 *tox* alleles. Among these, the most frequent are *tox*-2 (n=97, including the vaccine strain PW8), *tox*-3 (n=81)
479 and *tox*-1 (n=41). These alleles are synonymous and thus result in the same amino acid sequence of the
480 diphtheria toxin, implying complete match with the vaccine strain toxin. Alleles *tox*-24, 25, 26,27, 35, 36,
481 37 and 38 were predicted as NTTB by DIPHTOSCAN. The potential impact of protein changes deduced from
482 *tox* gene sequence variation was previously analyzed (Will et al., 2021); we provide the correspondence of
483 *tox* alleles in this previous study and ours in **Figure S8**.
484 The diversity of *tox*-positive isolates was evident from their distribution in the *C. diphtheriae*
485 phylogenetic tree, but it was striking that the Gravis branch comprised much less *tox*-positive sublineages
486 than the Mitis branch (**Figure 3**): in the Gravis lineage, there were only three main branches of *tox*-positive
487 isolates: (i) an early-branching group of sublineages; (ii) a branch comprising SL377 and SL384 (two frequent
488 sublineages in France-2022), and (iii) SL8. NTTB isolates were only observed in the Mitis lineage (with one
489 exception in Gravis-SL384) and this phenotype was acquired through multiple independent evolutionary
490 events (**Figure 3**).
491 A high diversity of *tox*-negative sublineages was also observed in the global dataset: whereas 173 of
492 245 (70.6%) sublineages were entirely *tox*-negative, only 73 (29.8%) of them had at least 1 *tox*-positive
493 isolate. Of these, 50 sublineages were homogeneous for *tox* status (*i.e.*, they included uniquely *tox*-positive
494 genomes), whereas 23 sublineages (9.3%) included both *tox*-positive and *tox*-negative genomes (**Table S1**;
495 **Figure 2**), indicating that the gain or loss of the *tox* gene is not uncommon within sublineages. When
496 considering the genomic clusters, almost all were either *tox*-positive or *tox*-negative in the global dataset.
497 Accordingly, sublineages in the France-2022 dataset were all either *tox* positive or negative, but notably,
498 SL377-GC71 comprised both types of isolates (**Figure 1**).
499
500 **6. Antimicrobial resistance**
501 DIPHTOSCAN includes a screen of *C. diphtheriae* genomes for the presence of antimicrobial resistance
502 genes or mutations against 10 classes of antimicrobial agents. DIPHTOSCAN also computes a resistance score,

503 defined as the number of antimicrobial classes for which at least one resistance gene or mutation is
504 detected. The resistance score varied from 0 to 8 in the global dataset; 38.2% non-redundant global isolates
505 had at least one genomic resistance feature, and 118 isolates (12.1%) were multidrug resistant (acquired
506 resistance to ≥3 drug classes; **Table S1**).

507 Resistance feature frequencies are shown in **Figure 4B** for the global dataset. The highest frequencies
508 of resistance genes were observed for sulfonamides (exclusively gene *sul1*; rarely present in two copies;
509 260 non-redundant isolates; 26.6%) and for tetracycline resistance, where *tet(O), tet(W)* and *tet(33)* were
510 present in approximately equal proportions (132 isolates; 13.5% in total). The phenicol resistance gene *cmx*
511 was also commonly found. *pbp2m* was present in 34 (3.5%) isolates, and *ermX* [sometimes named *erm(X)*]
512 in 36 (3.7%) isolates, with 14 (1.4%) isolates carrying both *pbp2m* and *ermX*.

513 Antimicrobial resistance genes were dispersed across the global *C. diphtheriae* phylogenetic tree
514 (**Figure 3**). The distribution of resistance at the sublineage level showed that just above half of the
515 sublineages (128; 52.0%) comprised at least one strain with at least one resistance genomic feature (**Table**
516 **S1**). The two sublineages with the most resistant strains were SL8 (the main sublineage involved in the ex-
517 USSR outbreak; 46 strains) and SL377 (17 strains) (**Figure 2**). 19 sublineages carried at least one multidrug
518 resistant isolate, and SL377 and SL405 were the most frequent of these (**Figure 2**).

519 Against this background, the France-2022 isolates appeared to carry resistance features much
520 more frequently, including *pbp2m, ermX* and quinolone-resistance determining mutations (**Figures 1 and**
521 **4**). 61 (60.4%) isolates presented at least one resistance feature (**Table S1**; **Figure 1**), and 44 (43.6%) were
522 multidrug resistant.

523 First-line treatments of diphtheria are penicillin or amoxicillin and macrolides in case of allergy to
524 beta-lactams. The *pbp2m* gene confers decreased susceptibility to penicillin and other beta-lactams (Forde
525 et al., 2020; Hennart et al., 2020), whereas *ermX* (and rarely *ermC*) are associated with erythromycin
526 resistance in *C. diphtheriae* (Tauch et al., 1995, 2003). In the global dataset, 34 isolates (**Table S1;** including
527 strain BQ11 with three copies consistent with Forde *et al*. 2020) carried *pbp2m* and 35 carried *ermX;* 14
528 (1.4%) isolates carried both genes. Sublineages SL297 and SL484 were the most common carriers of these
529 genes, whereas the frequent multidrug resistant sublineages SL377, SL384 and SL301 did not carry *ermX*
530 and *pbp2m* (**Figure S9**). In France-2022, 8 (7.9%) isolates carried both *pbp2m* and *ermX*. These were
531 observed in patients with travel history from Mali (SL395, SL542, SL852) and Egypt (SL297-GC820).

532 Antimicrobial susceptibility phenotypes were determined for the France-2022 dataset, and were
533 highly concordant with the presence of resistance features (**Table S4**). Resistance to penicillin and
534 macrolides was associated with *pbp2m* and *ermX*, respectively, although some *ermX*-carrying isolates
535 remained susceptible to erythromycin (**Table S4**).

536 We included in DIPHTOSCAN a search for integrons, which may harbor multiple resistance genes in
537 *C. diphtheriae* (Barraud et al., 2011; Arcari et al., 2023). In the global dataset, we identified 45 (4.6%) isolates
538 carrying integrons (including integrase-less ones, *i.e.*, CALINs) (**Table S1**), which were highly dispersed in the
539 phylogeny (not shown). In France-2022, we found the presence of complete integrons in 9 isolates and
540 integrase-less integrons in 9 additional isolates (18; 17.8%). These structures were strongly associated with
541 antimicrobial resistance, particularly to trimethoprim and sulfonamides (**Figure 1; Table S1**).

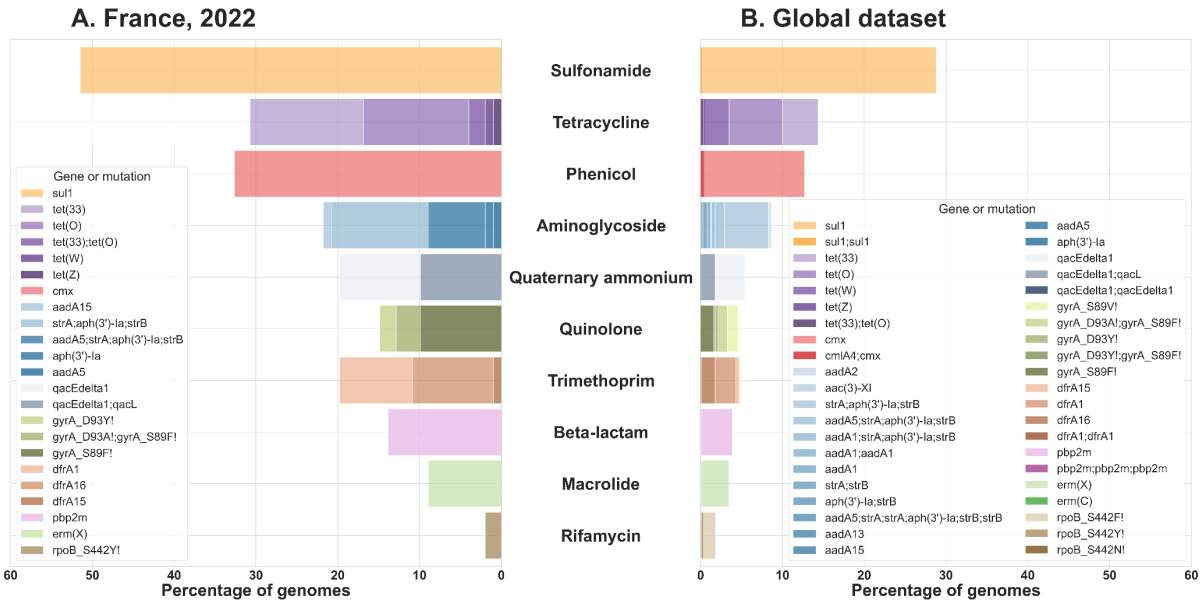

Figure 4. Observed frequencies of resistance genes or mutations

The number of genomes with a genetic feature associated with resistance, per antimicrobial class. Left: Isolates from France, 2022 (n=101 genomes); Right: global deduplicated dataset (n=976 genomes). The bars are ordered vertically by decreasing frequency in the right panel and the bar sectors are colored according to the presence of resistance features (see keys).

7. Dual risk isolates: convergence of diphtheria toxin and multidrug resistance, including to first-line treatments

The presence within the same isolates of multidrug resistance and toxigenicity could cause particularly threatening infections. We therefore explored the co-occurrence of these two genotypes (**Figure 2**). In the global dataset, 57 (5.8%) isolates were both multidrug resistant and *tox*-positive. The majority of these isolates belonged to a few sublineages (**Figure 2**), including SL377, which comprised 9 *tox*-positive multidrug resistant isolates mostly from India (and also observed in France-2022). Eight convergent isolates of SL301 were also observed from India, Austria and Syria. SL453 had three *tox*-positive multidrug resistant isolates, which were isolated in Spain and France with links to Afghanistan (Arcari et al., 2023). In metropolitan France, there were 22 *tox*-positive isolates that were multidrug resistant (21.8%), with SL377 and SL696 being predominant among these (**Table S1**, **Figure 1**).

Regarding resistance genes to first-line treatments, there was not a single isolate carrying at the same time *tox, pbp2m* and *ermX* in the global dataset (**Table S1**). However, in France-2022, SL852 isolates (from two patients with travel history from Mali) were *tox*-positive and carried *pbp2m* and *ermX*. Furthermore, they carried other resistance genes including *cmx*, *sul1, dfrA1,* and in addition *tet33* and *aadA15* for isolate FRC1688. This latter isolate only lacked resistance features to quinolones and rifampicin. No other isolate of this particularly concerning sublineage (SL852) was found in the global dataset.

8. Lineages Gravis and Mitis differ in the presence of pathogenicity-associated genes

Biovars represent an early attempt to discriminate among *C. diphtheriae* strains (Anderson et al., 1931) and are still commonly reported. We found that lineages Mitis and Gravis, defined genetically based on the

570 presence of the *spuA* gene probably involved in starch utilization, correspond to two distinct parts of the
571 phylogenetic tree (**Figure 3**) as previously reported (Hennart et al., 2020; Guglielmini et al., 2021). Note that
572 the match between lineage and *spuA* or biovar phenotype is not absolute, as a few isolates within the Gravis
573 branch were *spuA*-negative (in particular SL625, SL130, SL102, and SL377) and 42 (5.1%) isolates of the Mitis
574 lineage were *spuA*-positive. Among the France-2022 isolates, for which biovars were in addition determined
575 phenotypically, the two biovars were also phylogenetically distinct (**Figure 1**). Nearly four in five (n=78) of
576 the France-2022 isolates had a Mitis biotype (including 37 *tox*-positives), with 23 Gravis strains (8 *tox*-
577 positive).

578 To provide a population-level view of pathogenesis features in *C. diphtheriae*, we included in the
579 DIPHTOSCAN database of searched genes, in addition to the *tox* gene, all virulence genes previously
580 demonstrated or strongly suspected to be involved in diphtheria pathogenesis (see **Table S2** for
581 pathogenesis involvement evidence). These include genes involved in iron and heme acquisition, fimbriae
582 biosynthesis and assembly, and other adhesins (Ott et al., 2022).

583 Screening for these genes in the global dataset revealed highly heterogeneous patterns of presence
584 and phylogenetic distribution (**Table S1**; **Figure S10**). We found that a number of virulence factors are highly
585 conserved within *C. diphtheriae;* for example, DIP1546 was present in all genomes except in 28_DSM43988,
586 and DIP0733, DIP1281, DIP1621, and DIP1880 were fully conserved (**Table S1**). The corynebactin transport
587 (*ciuA-D*) gene cluster was present in all genomes, with one exception, whereas the corynebactin synthesis
588 (*ciuEFG*) locus was absent or incomplete in only 5.4% of genomes (n=29 Mitis, n=25 Gravis); of these, 33
589 lacked the *ciuE* gene, which is essential for siderophore synthesis. One of the genomes lacking *ciuE*
590 corresponds to the vaccine strain PW8, which is defective for corynebactin synthesis (Russell & Holmes,
591 1985). The heme-acquisition genes *hmuTUV* were also largely conserved (921 genomes; 94.4%).

592 In contrast, some genes were infrequent: DIP2014, a gene encoding for a BigA-like adhesin, was
593 detected in only a few sublineages of the Gravis branch (133 isolates), and the DIP0543 (also known as
594 *nanH*, coding for a sialidase) was present in only a few sublineages distributed across the phylogeny (not
595 shown).

596 Remarkably, we uncovered a sharp divide between lineages Gravis and Mitis in terms of iron
597 metabolism-associated genes, fimbriae gene clusters and other genes (**Figure S10**). The putative
598 siderophore synthesis and transport operon *irp2ABCDEFI-irp2JKLMN* was strongly associated with the Mitis
599 lineage: 513 out of 567 (90.5%) Mitis isolates were *irp2*-positive, whereas only 1 of 406 Gravis isolates was
600 *irp2*-positive. The iron transport cluster *irp1ABCD* was also mainly present in the Mitis lineage. Differently,
601 the *htaA* gene, which is part of the same gene cluster as *hmuTUV* and codes for a membrane protein that
602 binds hemoglobin, was absent or truncated in most genomes from the Mitis branch (92.1%), whereas it was
603 largely conserved in the Gravis branch (99.8% *htaA*-positive). Similar to *htaA*, genes *chtA* and *chtB*, which
604 have sequence and functional similarity to *htaA* and *htaB*, were also strongly associated with the Gravis
605 lineage: 304 of 406 Gravis isolates were *chtAB*-positive (74.9%), whereas only 7 of 567 Mitis isolates were
606 *chtAB*-positive (1.2%). In sharp contrast, the *htaC* gene, which is suspected to be involved in hemin
607 transport, and which is also in genetic linkage with the *hmuTUV* gene cluster, was entirely absent from the
608 Gravis branch, but was detected in 68.6% of Mitis genomes.

609 Three main fimbriae gene clusters, encoding fimbrial proteins, SpaA, SpaD and SpaH, have been
610 described in *C. diphtheriae* (Rogers et al., 2011; Reardon-Robinson & Ton-That, 2014; Sangal & Hoskisson,
611 2016). We found that these were more commonly found in the Gravis branch compared to the Mitis branch
612 (**Figure S10**). The SpaH gene cluster (*spaGHI-srtDE*) was present in its entirety in 254 genomes and as a
613 cluster with one missing gene in 29 isolates, all of which belonged to the Gravis lineage. The other two
614 systems showed some variability in the distribution of their genes. The sortase-mediated assembly genes
615 of the SpaA type pili, *spaABC,* were found in biovar Gravis in similar proportions (87.2% *spaA*, 86.2% *spaB*
616 and 86.0% *spaC*-positive), whereas in Mitis *spaB* was present in about half of the genomes (49.0%) and
617 *spaA* and *spaC* in one third (17.5%, and 18.2%, respectively). The distribution of the SpaA pilin-specific
618 sortase gene *srtA* was similar to that of *spaB* (98.8% in Gravis, 49.9% in Mitis), and the complete SpaA gene
619 cluster *spaABC-srtA* was found in only 299 genomes (30.6%), the majority of which were of Gravis lineage
620 (n=256). Last, genes of the SpaD cluster were less frequent (*spaD* 8.7%, *spaE* 14.9%, *spaF* 9.3%, *srtB* 33.2%,
621 *srtC* 33.7%) compared to the other pili types, and the complete gene cluster (*spaDEF-srtBC*) was found only
622 in 11 genomes, all of which belonged to lineage Gravis. Interestingly, the presence of SpaD and SpaH
623 complemented each other in the Gravis branch (**Figure S10**).
624 We further found that the collagen-binding protein DIP2093 (Peixoto et al.,2017) is strongly
625 associated with the Gravis lineage: 118 of 406 (29.1%) Gravis isolates were DIP2093-positive, whereas only
626 3 of 567 (0.5%) Mitis isolates were.
627 The complement of virulence genes of the France-2022 isolates was in full agreement with their
628 Gravis/Mitis placement and the above observations. For example, the *irp2A-I* and *irp2J-N* gene clusters
629 were present uniquely in sublineages belonging to the Mitis branch, and the *htaC* gene was present only in
630 64.2% of the Mitis genomes (**Table S1**); *chtA* and *chtB* were completely absent in Mitis and the collagen-
631 binding protein DIP2093 uniquely in Gravis isolates (n=16, 47.1%). None of the France-2022 isolates carried
632 a complete SpaD fimbriae cluster; in particular, they all lacked at least the *spaD* gene; and only 8 Gravis
633 genomes carried the complete SpaH cluster. The latter were dispersed among various lineages (SL32, SL374,
634 SL502, SL542, SL130).
635

636 ## Discussion

637 In recent years, large epidemics of diphtheria have been observed, *e.g.,* in South Africa, Bangladesh
638 and Yemen (du Plessis et al., 2017; Polonsky et al., 2021; Badell et al., 2021), while a progressive increase
639 of diphtheria cases has been noted in multiple countries (Bernard et al., 2019; Truelove et al., 2020).
640 However, so far, our understanding of diphtheria reemergence has been hindered by a lack of background
641 knowledge on the population diversity of *C. diphtheriae*, its sublineages of concern and the epidemiology
642 of their local or global dissemination. Here, we report on a sharp increase in *tox*-positive *C. diphtheriae* in
643 France in 2022, and developed a bioinformatics pipeline, DIPHTOSCAN, which enables to harmonize the way
644 genomic diversity and genetic features of medical concern are detected, reported and interpreted. We
645 illustrate how this novel tool provides clinically-relevant genomic profiling and evolutionary understanding

646 of emergence, by placing the 2022 *C. diphtheriae* from France in the context of 1,249 global *C. diphtheriae*
647 genomes.

648 Our results provide an updated overview of the population diversity of *C. diphtheriae* based on
649 currently available genomic sequences. As previously reported (Berger et al., 2019; Seth-Smith & Egli, 2019;
650 Hennart et al., 2020; Guglielmini et al., 2021), *C. diphtheriae* is made up of multiple sublineages that are
651 related through a star-like phylogeny. We here uncovered 123 novel sublineages, for a total of 253
652 described ones. We observed that, compared to previous datasets, there was no sublineage fusion upon
653 adding novel genomes, which indicated an excellent stability of *C. diphtheriae* sublineage classification. The
654 latter provides a broad classification of isolates that correlates strongly with classical MLST, and which
655 facilitates a deep-level approach to *C. diphtheriae* diversity and evolution. The naming of sublineages by
656 inheritance of ST numbers will facilitate continuity with classical MLST. Besides, sublineage classification is
657 more congruent with phylogenetic relationships: whereas most (140/146; 95.8%) non-singleton
658 sublineages were monophyletic, only 134 of 167 (79.8%) non-singleton STs were (data not shown). We
659 therefore strongly recommend transitioning from MLST to the cgMLST-based nomenclature, which is
660 available on the BIGSdb-Pasteur platform. Our phylogenetic analysis of reference strains of the historical
661 ribotype nomenclature provides a first overview of their relationships, to our knowledge, and allows
662 revisiting genealogical inferences that were made among ribotypes based on CRISPR spacer variation
663 (Mokrousov, 2009).

664 Genomic clusters represent a much narrower genetic classification of *C. diphtheriae* isolates,
665 compatible with recent transmission (Guglielmini et al., 2021). Therefore, genomic clusters appear more
666 relevant than sublineages for epidemiological investigation purposes, as illustrated for example within
667 SL377: whereas GC817 was associated with Afghanistan, GC71 was associated with Senegal and these two
668 genomic clusters of sublineage SL377 were clearly distinct phylogenetically (**Figure 1**).

669 The diagnostic and surveillance of diphtheria is largely based on the detection of the *tox* gene and
670 its expression (WHO, 2018). We found that the determination of the *tox* gene presence by DIPHTOSCAN was
671 highly concordant with the experimental reference qPCR. We also found that DIPHTOSCAN can predict a large
672 proportion of non-toxigenic *tox* gene-bearing (NTTB) isolates. Still, some NTTB isolates were not identified
673 by DIPHTOSCAN. These cases may be attributable to (i) a lack of detection by the Elek test due to a low level
674 of expression of the toxin gene in some strains, or (ii) yet unknown genetic mechanisms that abort *tox* gene
675 expression entirely (unexplained true NTTB). Future work is needed to define the genotype-phenotype links
676 underlying toxigenicity and to improve our predictive capacity of toxigenicity from genomic sequences. In
677 the non-redundant global dataset, 16.0% of *tox*-positive isolates were predicted as NTTB, which provides a
678 quantitative view of the relevance of differentiating mere *tox* gene presence from actual toxigenicity. The
679 capacity to predict toxigenicity from sequences opens interesting perspectives as to the diagnostic of
680 diphtheria based on rapid genomic sequencing. Our phylogenetic analysis showed that gain or loss of the
681 *tox* gene is a rare event at the timescale of genomic cluster diversification. The phenomenon of *tox* status
682 switch by phage acquisition or loss during infection or transmission was suspected
683 previously (Pappenheimer & Murphy, 1983) and deserves further study given its importance for public
684 health and clinical management.

685 Up until now, antimicrobial resistance has been considered of moderate clinical concern in
686 *C. diphtheriae* (Zasada, 2014; WHO, 2018). Although resistant strains have been described, clinical
687 susceptibility breakpoints have lacked standardization and the prevalence, origin and dissemination of
688 resistance genetic features are largely unknown. Here, we identified in the France-2022 isolates as well as
689 in the global *C. diphtheriae*, multidrug resistant isolates and/or isolates resistant to first-line treatments.
690 We provide an overview of the prevalence and distribution of resistance genes or mutations in
691 *C. diphtheriae*, and identify sublineages that carry multiple resistance genes. Because antimicrobial
692 resistance phenotypes are typically unattached to publicly available genomic sequences, it is not possible
693 to link these genomic features complements to resistance phenotypes systematically. However, this (**Table**
694 **S4**) and previous works clearly showed that most resistance genetic features identified here may impact
695 resistance phenotypes (Tauch et al., 1995, 2003; Hennart et al., 2020; Forde et al., 2020). Of particular
696 concern, *tox*-positive isolates that are resistant to multiple drugs and/or first-line treatments were
697 identified herein, with the convergence of *tox*, *pbp2m* an *ermX* in two 2022 cases with a travel history from
698 Mali, which were resistant to 9 and 11 out of 23 tested antimicrobials, respectively. Such isolates may pose
699 serious clinical management difficulties, and multidrug resistant *C. diphtheriae* should therefore be closely
700 monitored.
701 The combined analysis of the France-2022 and global datasets using a unique pipeline provides context
702 to the reemergence of diphtheria (**Figure S6**). The occurrence of cases of diphtheria among migrants, the
703 vast majority of whom are not up to date with their vaccinations, raises concerns of the emergence of
704 cluster cases in accommodation facilities for migrants, refugees or asylum seekers (Badenschier et al., 2022;
705 Kofler et al., 2022). Professionals dealing with these populations need to be particularly vigilant in spotting
706 clinical signs of diphtheria and ensuring that their vaccinations are up to date. Here, we found that some
707 sublineages contributing to the reemergence were previously observed, whereas others are described for
708 the first time. For example, SL377, one of the major toxigenic and resistant sublineages observed in France-
709 2022, had been circulating in India during 2016 and was reported in Europe (Spain and France) since 2015
710 (**Table S1**). In contrast, SL698 was absent from the global dataset. Of the 10 *tox*-positive France-2022
711 sublineages, five were associated with travel from Afghanistan, and were recently described in other
712 European countries too (Badenschier et al., 2022; Kofler et al., 2022).
713 The DIPHTOSCAN tool will facilitate the harmonized characterization of *C. diphtheriae* sublineages of
714 concern. Several virulence-associated genes were largely conserved in the entire *C. diphtheriae* population
715 analyzed; these genomic features may therefore be central for *C. diphtheriae* colonization and transmission
716 among humans, as there appears to be a strong selective pressure to maintain them. The distribution of
717 other, more variably present, virulence-associated genes uncovers a very striking dichotomy between the
718 Gravis and Mitis lineages, as heme and iron-acquisition systems and Spa-encoded fimbriae gene clusters
719 were either associated with the Mitis or the Gravis lineages, in a largely mutually exclusive way. Based on
720 these observations, the Gravis lineage may preferentially capture iron from hemin, whereas the Mitis one
721 could be associated with the ability to synthesize and use siderophores. There might be important
722 implications for the regulation and expression level of the *tox* gene, which is controlled by the iron-
723 dependent DtxR repressor. Importantly, the toxin gene and its NTTB-leading disruptions were also

724 unequally distributed between Gravis and Mitis lineages. It was noted early that toxin production is less
725 inhibited by infection-relevant iron concentrations in Gravis strains (Mueller, 1941; McLeod, 1943), and our
726 results shed a new light and provides experimentally testable hypotheses on this critical difference in the
727 biology of infection of the Gravis and Mitis lineages.

728 Another striking feature we uncovered is the distribution of gene clusters coding for fimbriae.
729 Previous work reported SpaA as being largely conserved in *C. diphtheriae*, with SpaD and SpaH being more
730 variably present (Reardon-Robinson & Ton-That, 2014; Sangal & Hoskisson, 2016; Ott, 2018). We found that
731 SpaA was largely present in our dataset, however, the complete gene cluster *spaABC-srtA* was mostly found
732 in the Gravis branch. SpaD was also more common among Gravis genomes, although the complete cluster
733 (*spaDEF-srtBC*) was only detected in a minority of genomes. None of the Mitis isolates were positive for
734 SpaH. These three Spa systems were experimentally shown to be involved in adhesion to different human
735 tissues: pharyngeal (SpaA), laryngeal (SpaD) and pulmonary (SpaH) epithelial cells (Mandlik et al., 2007;
736 Reardon-Robinson & Ton-That, 2014). The Gravis/Mitis dichotomy in Spa-type fimbriae may have important
737 implications regarding a possible differential ecology, transmission, tissue tropism and pathophysiology of
738 these two major *C. diphtheriae* lineages.

739 In conclusion, we developed and applied to a large dataset, the bioinformatics tool DIPHTOSCAN. Its
740 public availability and ease of use will enable to conveniently extract and interpret genomic features that
741 are relevant to the clinical and public health management of diphtheria cases, to understand the
742 microbiological determinants of (re)emerging sublineages, and to future research on the genotype-clinical
743 phenotype links in *C. diphtheriae*. This dedicated tool is also applicable to the other members of the
744 *C. diphtheriae* complex, such as *C. ulcerans* (data not shown). Harmonization of genomic studies in this
745 group of pathogens, which have been largely forgotten but currently undergo re-emergence in Europe and
746 elsewhere, will support genomic surveillance of diphtheria, will contribute to enhance our understanding
747 of the pathogenesis of modern diphtheria, and opens interesting hypotheses as to the underlying
748 mechanisms of variation in clinical severity and forms of diphtheria.

749

750 **Acknowledgements**

751 We thank Martin Maiden and Keith Jolley (Oxford University) for maintaining the previous MLST
752 data from Oxford's PubMLST database and for providing the data for import into the BIGSdb-Pasteur
753 *C. diphtheriae* species complex database.

754

755 **Funding**

756 MH was supported financially by the PhD grant "Codes4strains" from the European Joint
757 Programme One Health, which has received funding from the European Union's Horizon 2020 Research and
758 Innovation Programme under Grant Agreement No. 773830. This work used the computational and storage
759 services provided by the IT department at Institut Pasteur. The National Reference Center for
760 Corynebacteria of the Diphtheriae Complex is supported financially by the Ministry of Health (Public Health
761 France) and Institut Pasteur.

762 **Conflict of interest disclosure**

763 The authors declare no conflict of interest.

764

765 ## Author contributions

766 S. Brisse (S.B.) conceived, designed, and coordinated the study. Melanie Hennart (M.H.) developed
767 the DIPHTOSCAN tool with input from SB. M.H. and S.B. analyzed the genomic data. M.H. created the figures
768 and tables. S.B. and M.H. created the first draft of the manuscript, worked together to improve it and
769 reviewed the final version. Chiara Crestani analyzed the iron metabolism and fimbriae genes distribution
770 and wrote the first version of the corresponding sections. Sebastien Bridel performed the merger of the
771 Oxford PubMLST and BIGSdb-Pasteur databases. Annick Carmi-Leroy, Sylvie Brémont, Annie Landier,
772 Nathalie Armatys and Virginie Passet provided technical assistance with the microbiological
773 characterization and sequencing of the *C. diphtheriae* isolates. Edgar Badell and Julie Toubiana contributed
774 to the NRC operations coordination. Laure Fonteneau and Sophie Vaux coordinated diphtheria
775 epidemiological surveillance in France. All authors reviewed and approved the final contents of the
776 manuscript.

777

778 ## Data, scripts, code, and supplementary information availability

779 The latest version of the DIPHTOSCAN code will be available at
780 https://gitlab.pasteur.fr/BEBP/diphtoscan and the version used in this work in available at:
781 https://zenodo.org/record/7774709.

782 The genome sequence data generated in this work has been made publicly available through
783 NCBI/ENA bioproject PRJEB22103 (https://www.ebi.ac.uk/ena/browser/view/PRJEB22103).

784 The trees are available at https://itol.embl.de/shared/Pasteur_BEBP in the projet: 'Hennart et al.,
785 2023: diphtOscan'.

786 The supplementary appendix is available in zenodo at: https://doi.org/10.5281/zenodo.8123234

787

788 **Ethical approval statement**: Diphtheria is a notifiable disease in France. Phenotypic and genotypic
789 analyses of bacterial isolates were carried out within the framework of the mandate given to the National
790 Reference Center for Corynebacteria of the Diphtheriae Complex by the Ministry of Health (Public Health
791 France). All French bacteriological samples and data were collected in the frame of the French national
792 diphtheria surveillance and are collected, coded, shipped, managed and analyzed according to the French
793 National Reference Center protocols. Other strains were obtained from culture collections.

794

795 **Authors' license statement**: This research was funded, in whole or in part, by Institut Pasteur and
796 by European Union's Horizon 2020 research and innovation programme. For the purpose of open access,
797 the authors have applied a CC-BY public copyright license to any Author Manuscript version arising from
798 this submission.

799

800 **References**

801 Anderson JS, Happold FC, McLeod JW, Thomson JG (1931) On the existence of two forms of diphtheria
802 bacillus—B. Diphtheriæ gravis and B. Diphtheriæ mitis—and a new medium for their
803 differentiation and for the bacteriological diagnosis of diphtheria. *The Journal of Pathology*
804 *and Bacteriology*, **34**, 667–681. https://doi.org/10.1002/path.1700340506

805 Arcari G, Hennart M, Badell E, Brisse S (2023) Multidrug-resistant toxigenic Corynebacterium
806 diphtheriae sublineage 453 with two novel resistance genomic islands. *Microbial Genomics*, **9**.
807 https://doi.org/10.1099/mgen.0.000923

808 Badell E, Alharazi A, Criscuolo A, Almoayed KAA, Lefrancq N, Bouchez V, Guglielmini J, Hennart M,
809 Carmi-Leroy A, Zidane N, Pascal-Perrigault M, Lebreton M, Martini H, Salje H, Toubiana J,
810 Dureab F, Dhabaan G, Brisse S, Rawah AA, Aldawla MA, Al-Awdi EM, Al-Moalmy NM, Al-Shami
811 HZ, Al-Somainy AA (2021) Ongoing diphtheria outbreak in Yemen: a cross-sectional and
812 genomic epidemiology study. *The Lancet Microbe*, **2**, e386–e396.
813 https://doi.org/10.1016/S2666-5247(21)00094-X

814 Badell E, Guillot S, Tulliez M, Pascal M, Panunzi LG, Rose S, Litt D, Fry NK, Brisse S (2019) Improved
815 quadruplex real-time PCR assay for the diagnosis of diphtheria. *Journal of Medical*
816 *Microbiology*, **68**, 1455–1465. https://doi.org/10.1099/jmm.0.001070

817 Badell E, Hennart M, Rodrigues C, Passet V, Dazas M, Panunzi L, Bouchez V, Carmi-Leroy A, Toubiana J,
818 Brisse S (2020) *Corynebacterium rouxii* sp. nov., a novel member of the diphtheriae species
819 complex. *Research in Microbiology*. https://doi.org/10.1016/j.resmic.2020.02.003

820 Badenschier F, Berger A, Dangel A, Sprenger A, Hobmaier B, Sievers C, Prins H, Dörre A, Wagner-
821 Wiening C, Külper-Schiek W, Wichmann O, Sing A (2022) Outbreak of imported diphtheria with
822 Corynebacterium diphtheriae among migrants arriving in Germany, 2022. *Euro Surveillance:*
823 *Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*,
824 **27**, 2200849. https://doi.org/10.2807/1560-7917.ES.2022.27.46.2200849

825 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S,
826 Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012)
827 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.
828 *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **19**, 455–
829 477. https://doi.org/10.1089/cmb.2012.0021

830 Barksdale L (1970) Corynebacterium diphtheriae and its relatives. *Bacteriological Reviews*, **34**, 378–
831 422.

832 Barraud O, Badell E, Denis F, Guiso N, Ploy M-C (2011) Antimicrobial drug resistance in
833 Corynebacterium diphtheriae mitis. *Emerging Infectious Diseases*, **17**, 2078–2080.
834 https://doi.org/10.3201/eid1711.110282

835 Benamrouche N, Hasnaoui S, Badell E, Guettou B, Lazri M, Guiso N, Rahal K (2016) Microbiological and
836 molecular characterization of Corynebacterium diphtheriae isolated in Algeria between 1992

837 and 2015. *Clinical Microbiology and Infection: The Official Publication of the European Society*
838 *of Clinical Microbiology and Infectious Diseases*, **22**, 1005.e1-1005.e7.
839 https://doi.org/10.1016/j.cmi.2016.08.013
840 Berger A, Dangel A, Schober T, Schmidbauer B, Konrad R, Marosevic D, Schubert S, Hörmansdorfer S,
841 Ackermann N, Hübner J, Sing A (2019) Whole genome sequencing suggests transmission of
842 Corynebacterium diphtheriae-caused cutaneous diphtheria in two siblings, Germany, 2018.
843 *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European*
844 *Communicable Disease Bulletin*, **24**. https://doi.org/10.2807/1560-
845 7917.ES.2019.24.2.1800683
846 Bernard KA, Pacheco AL, Burdz T, Wiebe D (2019) Increase in detection of Corynebacterium
847 diphtheriae in Canada: 2006-2019. *Canada Communicable Disease Report = Releve Des*
848 *Maladies Transmissibles Au Canada*, **45**, 296–301. https://doi.org/10.14745/ccdr.v45i11a04
849 Bolt F, Cassiday P, Tondella ML, Dezoysa A, Efstratiou A, Sing A, Zasada A, Bernard K, Guiso N, Badell E,
850 Rosso ML, Baldwin A, Dowson C (2010) Multilocus sequence typing identifies evidence for
851 recombination and two distinct lineages of Corynebacterium diphtheriae. *J Clin Microbiol*, **48**,
852 4177–85. https://doi.org/10.1128/JCM.00274-10
853 Bonmarin I, Guiso N, Le Flèche-Matéos A, Patey O, Grimont Patrick AD, Levy-Bruhl D (2009) Diphtheria:
854 A zoonotic disease in France? *Vaccine*, **27**, 4196–4200.
855 https://doi.org/10.1016/j.vaccine.2009.04.048
856 Chorlton SD, Ritchie G, Lawson T, Romney MG, Lowe CF (2019) Whole-genome sequencing of
857 Corynebacterium diphtheriae isolates recovered from an inner-city population demonstrates
858 the predominance of a single molecular strain. *Journal of Clinical Microbiology*, **58**, e01651-19.
859 https://doi.org/10.1128/JCM.01651-19
860 Criscuolo A (2020) On the transformation of MinHash-based uncorrected distances into proper
861 evolutionary distances for phylogenetic inference. *F1000Research*, **9**, 1309.
862 https://doi.org/10.12688/f1000research.26930.1
863 Criscuolo A, Brisse S (2013) AlienTrimmer: A tool to quickly and accurately trim off multiple short
864 contaminant sequences from high-throughput sequencing reads. *Genomics*,
865 10.1016/j.ygeno.2013.07.011. https://doi.org/10.1016/j.ygeno.2013.07.011
866 Cury J, Jové T, Touchon M, Néron B, Rocha EP (2016) Identification and analysis of integrons and
867 cassette arrays in bacterial genomes. *Nucleic Acids Research*, **44**, 4539–4550.
868 https://doi.org/10.1093/nar/gkw319
869 Dangel A, Berger A, Konrad R, Bischoff H, Sing A (2018) Geographically Diverse Clusters of Nontoxigenic
870 Corynebacterium diphtheriae Infection, Germany, 2016-2017. *Emerging Infectious Diseases*,
871 **24**, 1239–1245. https://doi.org/10.3201/eid2407.172026

872 Dangel A, Berger A, Konrad R, Sing A (2019) NGS-based phylogeny of diphtheria-related pathogenicity
873 factors in different Corynebacterium spp. implies species-specific virulence transmission. *BMC*
874 *microbiology*, **19**, 28. https://doi.org/10.1186/s12866-019-1402-1
875 Dangel A, Berger A, Rau J, Eisenberg T, Kämpfer P, Margos G, Contzen M, Busse H-J, Konrad R, Peters
876 M, Sting R, Sing A (2020) Corynebacterium silvaticum sp. nov., a unique group of NTTB
877 corynebacteria in wild boar and roe deer. *International Journal of Systematic and Evolutionary*
878 *Microbiology*, **70**, 3614–3624. https://doi.org/10.1099/ijsem.0.004195
879 Dazas M, Badell E, Carmi-Leroy A, Criscuolo A, Brisse S (2018) Taxonomic status of Corynebacterium
880 diphtheriae biovar Belfanti and proposal of Corynebacterium belfantii sp. nov. *International*
881 *Journal of Systematic and Evolutionary Microbiology*, **68**, 3826–3831.
882 https://doi.org/10.1099/ijsem.0.003069
883 Engler KH, Glushkevich T, Mazurova IK, George RC, Efstratiou A (1997) A modified Elek test for
884 detection of toxigenic corynebacteria in the diagnostic laboratory. *Journal of Clinical*
885 *Microbiology*, **35**, 495–498.
886 Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill
887 JB, Prasad AB, Tillman GE, Tyson GH, Klimke W (2021) AMRFinderPlus and the Reference Gene
888 Catalog facilitate examination of the genomic links among antimicrobial resistance, stress
889 response, and virulence. *Scientific Reports*, **11**, 12728. https://doi.org/10.1038/s41598-021-
890 91456-0
891 Forde BM, Henderson A, Playford EG, Looke D, Henderson BC, Watson C, Steen JA, Sidjabat HE, Laurie
892 G, Muttaiyah S, Nimmo GR, Lampe G, Smith H, Jennison AV, McCall B, Carroll H, Cooper MA,
893 Paterson DL, Beatson SA (2020) Fatal respiratory diphtheria caused by β-lactam-resistant
894 Corynebacterium diphtheriae. *Clinical Infectious Diseases*, **73**, e4531–e4538.
895 Grimont PAD, Grimont F, Efstratiou A, De Zoysa A, Mazurova I, Ruckly C, Lejay-Collin M, Martin-
896 Delautre S, Regnault B, European Laboratory Working Group on Diphtheria (2004)
897 International nomenclature for Corynebacterium diphtheriae ribotypes. *Research in*
898 *Microbiology*, **155**, 162–166. https://doi.org/10.1016/j.resmic.2003.12.005
899 Guglielmini J, Hennart M, Badell E, Toubiana J, Criscuolo A, Brisse S (2021) Genomic Epidemiology and
900 Strain Taxonomy of Corynebacterium diphtheriae. *Journal of Clinical Microbiology*, **59**,
901 e0158121. https://doi.org/10.1128/JCM.01581-21
902 Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, Criscuolo A, Brisse S (2022) A Dual Barcoding
903 Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of Klebsiella pneumoniae
904 Strains. *Molecular Biology and Evolution*, **39**, msac135.
905 https://doi.org/10.1093/molbev/msac135
906 Hennart M, Panunzi LG, Rodrigues C, Gaday Q, Baines SL, Barros-Pinkelnig M, Carmi-Leroy A, Dazas M,
907 Wehenkel AM, Didelot X, Toubiana J, Badell E, Brisse S (2020) Population genomics and

908 antimicrobial resistance in Corynebacterium diphtheriae. *Genome Medicine*, **12**, 107.
909 https://doi.org/10.1186/s13073-020-00805-7
910 Hoefer A, Pampaka D, Herrera-León S, Peiró S, Varona S, López-Perea N, Masa-Calles J, Herrera-León L
911 (2020) Molecular and epidemiological characterisation of toxigenic and non-toxigenic C.
912 diphtheriae, C. belfantii and C. ulcerans isolates identified in Spain from 2014 to 2019. *Journal*
913 *of Clinical Microbiology*, **59**, e02410-20. https://doi.org/10.1128/JCM.02410-20
914 Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S (2018) High throughput ANI analysis
915 of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, **9**,
916 5114. https://doi.org/10.1038/s41467-018-07641-9
917 Jolley KA, Maiden MC (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population
918 level. *BMC Bioinformatics*, **11**, 595. https://doi.org/10.1186/1471-2105-11-595
919 Kofler J, Ramette A, Iseli P, Stauber L, Fichtner J, Droz S, Zihler Berner A, Meier AB, Begert M, Negri S,
920 Jachmann A, Keller PM, Staehelin C, Grützmacher B (2022) Ongoing toxin-positive diphtheria
921 outbreaks in a federal asylum centre in Switzerland, analysis July to September 2022. *Euro*
922 *Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable*
923 *Disease Bulletin*, **27**, 2200811. https://doi.org/10.2807/1560-7917.ES.2022.27.44.2200811
924 Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for
925 prokaryotes. *Proceedings of the National Academy of Sciences*, **102**, 2567–2572.
926 https://doi.org/10.1073/pnas.0409727102
927 Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE (2021) A genomic surveillance framework
928 and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nature*
929 *Communications*, **12**, 4188. https://doi.org/10.1038/s41467-021-24448-3
930 Liu Y, Schröder J, Schmidt B (2013) Musket: a multistage k-mer spectrum-based error corrector for
931 Illumina sequence data. *Bioinformatics (Oxford, England)*, **29**, 308–315.
932 https://doi.org/10.1093/bioinformatics/bts690
933 Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome
934 assemblies. *Bioinformatics*, **27**, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507
935 Mandlik A, Swierczynski A, Das A, Ton-That H (2007) Corynebacterium diphtheriae employs specific
936 minor pilins to target human pharyngeal epithelial cells. *Molecular microbiology*, **64**, 111–124.
937 https://doi.org/10.1111/j.1365-2958.2007.05630.x
938 McLeod JW (1943) THE TYPES MITIS, INTERMEDIUS AND GRAVIS OF CORYNEBACTERIUM
939 DIPHTHERIAE: A Review of Observations during the Past Ten Years. *Bacteriological Reviews*, **7**,
940 1–41.
941 Meinel DM, Kuehl R, Zbinden R, Boskova V, Garzoni C, Fadini D, Dolina M, Blümel B, Weibel T, Tschudin-
942 Sutter S, Widmer AF, Bielicki JA, Dierig A, Heininger U, Konrad R, Berger A, Hinic V,
943 Goldenberger D, Blaich A, Stadler T, Battegay M, Sing A, Egli A (2016) Outbreak investigation
944 for toxigenic Corynebacterium diphtheriae wound infections in refugees from Northeast Africa

945 and Syria in Switzerland and Germany by whole genome sequencing. *Clinical Microbiology and*
946 *Infection: The Official Publication of the European Society of Clinical Microbiology and*
947 *Infectious Diseases*, **22**, 1003.e1-1003.e8. https://doi.org/10.1016/j.cmi.2016.08.010
948 Melnikov VG, Berger A, Sing A (2022) Detection of diphtheria toxin production by toxigenic
949 corynebacteria using an optimized Elek test. *Infection*, **50**, 1591–1595.
950 https://doi.org/10.1007/s15010-022-01903-x
951 Mina NV, Burdz T, Wiebe D, Rai JS, Rahim T, Shing F, Hoang L, Bernard K (2011) Canada's first case of
952 a multidrug-resistant Corynebacterium diphtheriae strain, isolated from a skin abscess. *Journal*
953 *of Clinical Microbiology*, **49**, 4003–4005. https://doi.org/10.1128/JCM.05296-11
954 Mokrousov I (2009) Corynebacterium diphtheriae: genome diversity, population structure and
955 genotyping perspectives. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology*
956 *and Evolutionary Genetics in Infectious Diseases*, **9**, 1–15.
957 https://doi.org/10.1016/j.meegid.2008.09.011
958 Mueller JH (1941) Toxin-production as related to the clinical severity of diphtheria. , **42**, 353–360.
959 Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha EPC (2022) IntegronFinder 2.0: Identification
960 and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella.
961 *Microorganisms*, **10**, 700. https://doi.org/10.3390/microorganisms10040700
962 Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016) Mash: fast
963 genome and metagenome distance estimation using MinHash. *Genome Biology*, **17**, 132.
964 https://doi.org/10.1186/s13059-016-0997-x
965 Ott L (2018) Adhesion properties of toxigenic corynebacteria. *AIMS Microbiology*, **4**, 85–103.
966 https://doi.org/10.3934/microbiol.2018.1.85
967 Ott L, Möller J, Burkovski A (2022) Interactions between the Re-Emerging Pathogen Corynebacterium
968 diphtheriae and Host Cells. *International Journal of Molecular Sciences*, **23**, 3298.
969 https://doi.org/10.3390/ijms23063298
970 Pappenheimer AM, Murphy JR (1983) Studies on the molecular epidemiology of diphtheria. *Lancet*
971 *(London, England)*, **2**, 923–926. https://doi.org/10.1016/s0140-6736(83)90449-x
972 Peixoto RS, Antunes CA, Lourêdo LS, Viana VG, Santos CS dos, Fuentes Ribeiro da Silva J, Hirata Jr. R,
973 Hacker E, Mattos-Guaraldi AL, Burkovski A 2017 Functional characterization of the collagen-
974 binding protein DIP2093 and its influence on host–pathogen interaction and arthritogenic
975 potential of Corynebacterium diphtheriae. *Microbiology*, **163**, 692–701.
976 https://doi.org/10.1099/mic.0.000467
977 du Plessis M, Wolter N, Allam M, de Gouveia L, Moosa F, Ntshoe G, Blumberg L, Cohen C, Smith M,
978 Mutevedzi P, Thomas J, Horne V, Moodley P, Archary M, Mahabeer Y, Mahomed S, Kuhn W,
979 Mlisana K, McCarthy K, von Gottberg A (2017) Molecular Characterization of Corynebacterium
980 diphtheriae Outbreak Isolates, South Africa, March-June 2015. *Emerging Infectious Diseases*,
981 **23**, 1308–1315. https://doi.org/10.3201/eid2308.162039

982 Polonsky JA, Ivey M, Mazhar MKA, Rahman Z, le Polain de Waroux O, Karo B, Jalava K, Vong S, Baidjoe
983 A, Diaz J, Finger F, Habib ZH, Halder CE, Haskew C, Kaiser L, Khan AS, Sangal L, Shirin T, Zaki QA,
984 Salam MA, White K (2021) Epidemiological, clinical, and public health response characteristics
985 of a large outbreak of diphtheria among the Rohingya population in Cox's Bazar, Bangladesh,
986 2017 to 2019: A retrospective study. *PLoS medicine*, **18**, e1003587.
987 https://doi.org/10.1371/journal.pmed.1003587
988 Reardon-Robinson ME, Ton-That H (2014) Assembly and function of Corynebacterium diphtheriae pili.
989 In: *Corynebacterium diphtheriae and Related Toxigenic Species* , pp. 123–141. Springer,
990 Heidelberg.
991 Rogers EA, Das A, Ton-That H (2011) Adhesion by pathogenic corynebacteria. *Advances in Experimental*
992 *Medicine and Biology*, **715**, 91–103. https://doi.org/10.1007/978-94-007-0940-9_6
993 Russell LM, Holmes RK (1985) Highly toxinogenic but avirulent Park-Williams 8 strain of
994 Corynebacterium diphtheriae does not produce siderophore. *Infection and Immunity*, **47**, 575–
995 578. https://doi.org/10.1128/iai.47.2.575-578.1985
996 Sangal V, Burkovski A, Hunt AC, Edwards B, Blom J, Hoskisson PA (2014) A lack of genetic basis for
997 biovar differentiation in clinically important Corynebacterium diphtheriae from whole genome
998 sequencing. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and*
999 *Evolutionary Genetics in Infectious Diseases*, **21**, 54–57.
1000 https://doi.org/10.1016/j.meegid.2013.10.019
1001 Sangal V, Hoskisson PA (2016) Evolution, epidemiology and diversity of Corynebacterium diphtheriae:
1002 New perspectives on an old foe. *Infection, Genetics and Evolution: Journal of Molecular*
1003 *Epidemiology and Evolutionary Genetics in Infectious Diseases*, **43**, 364–370.
1004 https://doi.org/10.1016/j.meegid.2016.06.024
1005 Santos AS, Ramos RT, Silva A, Hirata R, Mattos-Guaraldi AL, Meyer R, Azevedo V, Felicori L, Pacheco
1006 LGC (2018) Searching whole genome sequences for biochemical identification features of
1007 emerging and reemerging pathogenic Corynebacterium species. *Functional & Integrative*
1008 *Genomics*, **18**, 593–610. https://doi.org/10.1007/s10142-018-0610-3
1009 Schaeffer J, Huhulescu S, Stoeger A, Allerberger F, Ruppitsch W (2020) Assessing the Genetic Diversity
1010 of Austrian Corynebacterium diphtheriae Clinical Isolates, 2011-2019. *Journal of Clinical*
1011 *Microbiology*. https://doi.org/10.1128/JCM.02529-20
1012 Seth-Smith HMB, Egli A (2019) Whole Genome Sequencing for Surveillance of Diphtheria in Low
1013 Incidence Settings. *Frontiers in Public Health*, **7**, 235.
1014 https://doi.org/10.3389/fpubh.2019.00235
1015 Tauch A, Bischoff N, Brune I, Kalinowski J (2003) Insights into the genetic organization of the
1016 Corynebacterium diphtheriae erythromycin resistance plasmid pNG2 deduced from its
1017 complete nucleotide sequence. *Plasmid*, **49**, 63–74. https://doi.org/10.1016/s0147-
1018 619x(02)00115-4

1019 Tauch A, Kassing F, Kalinowski J, Pühler A (1995) The Corynebacterium xerosis composite transposon
1020 Tn5432 consists of two identical insertion sequences, designated IS1249, flanking the
1021 erythromycin resistance gene ermCX. *Plasmid*, **34**, 119–131.
1022 https://doi.org/10.1006/plas.1995.9995
1023 Timms VJ, Nguyen T, Crighton T, Yuen M, Sintchenko V (2018) Genome-wide comparison of
1024 Corynebacterium diphtheriae isolates from Australia identifies differences in the Pan-genomes
1025 between respiratory and cutaneous strains. *BMC genomics*, **19**, 869, 1–10.
1026 https://doi.org/10.1186/s12864-018-5147-2
1027 Truelove SA, Keegan LT, Moss WJ, Chaisson LH, Macher E, Azman AS, Lessler J (2020) Clinical and
1028 Epidemiological Aspects of Diphtheria: A Systematic Review and Pooled Analysis. *Clinical*
1029 *Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, **71**,
1030 89–97. https://doi.org/10.1093/cid/ciz808
1031 WHO (2018) Diphtheria: Vaccine Preventable Diseases Surveillance Standards.
1032 https://www.who.int/publications/m/item/vaccine-preventable-diseases-surveillance-
1033 standards-diphtheria.
1034 Will RC, Ramamurthy T, Sharma NC, Veeraraghavan B, Sangal L, Haldar P, Pragasam AK, Vasudevan K,
1035 Kumar D, Das B, Heinz E, Melnikov V, Baker S, Sangal V, Dougan G, Mutreja A (2021)
1036 Spatiotemporal persistence of multiple, diverse clades and toxins of Corynebacterium
1037 diphtheriae. *Nature Communications*, **12**, 1500, 1–8. https://doi.org/10.1038/s41467-021-
1038 21870-5
1039 Williams MM, Waller JL, Aneke JS, Weigand MR, Diaz MH, Bowden KE, Simon AK, Peng Y, Xiaoli L,
1040 Cassiday PK, Winchell J, Tondella ML (2020) Detection and Characterization of Diphtheria Toxin
1041 Gene-Bearing *Corynebacterium* Species through a New Real-Time PCR Assay (DJ Diekema, Ed,).
1042 *Journal of Clinical Microbiology*, **58**. https://doi.org/10.1128/JCM.00639-20
1043 Xiaoli L, Benoliel E, Peng Y, Aneke J, Cassiday PK, Kay M, McKeirnan S, Duchin JS, Kawakami V, Lindquist
1044 S, Acosta AM, DeBolt C, Tondella ML, Weigand MR (2020) Genomic epidemiology of
1045 nontoxigenic Corynebacterium diphtheriae from King County, Washington State, USA between
1046 July 2018 and May 2019. *Microbial Genomics*, **6**. https://doi.org/10.1099/mgen.0.000467
1047 Zakikhany K, Neal S, Efstratiou A (2014) Emergence and molecular characterisation of non-toxigenic
1048 tox gene-bearing Corynebacterium diphtheriae biovar mitis in the United Kingdom, 2003-
1049 2012. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European*
1050 *Communicable Disease Bulletin*, **19**. https://doi.org/10.2807/1560-7917.es2014.19.22.20819
1051 Zasada AA (2014) Antimicrobial Susceptibility and Treatment. In: *Corynebacterium diphtheriae and*
1052 *Related Toxigenic Species: Genomics, Pathogenicity and Applications* (ed Burkovski A), pp.
1053 239–246. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-007-7624-1_12

1054