Supplementary appendix

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

Contents:

Supplementary tables S1 to S4

Supplementary figures S1 to S9

30

Supplementary tables

- *See separately attached tables -*

A. *tox+ C. diphtheriae* cases, France, 2002-2022

B. *tox+ C. diphtheriae* cases, France, 2022

68
69
70 **Figure S1. Epidemiological curve of *tox*-positive *C. diphtheriae* infections in France**

71 **A**. Number of reported cases by year, France, 2002-2022; **B**. Number of reported cases by week,

72 France, 2022. The cases were either from metropolitan France, or overseas territories (see color

73 key).

74
75
76
77
78
79
80

DIPHTOSCAN main analytical steps

1 → 2 → 3 → 4 → 5 → 6

if --tree option

if --integron option

get species | get virulence genes | get resistance genes | get biovar loci | get integrons | get tree

if species is in CdSC* | if *tox* is detected

get MLST and ST | get prediction of toxigenicity | get genomic context
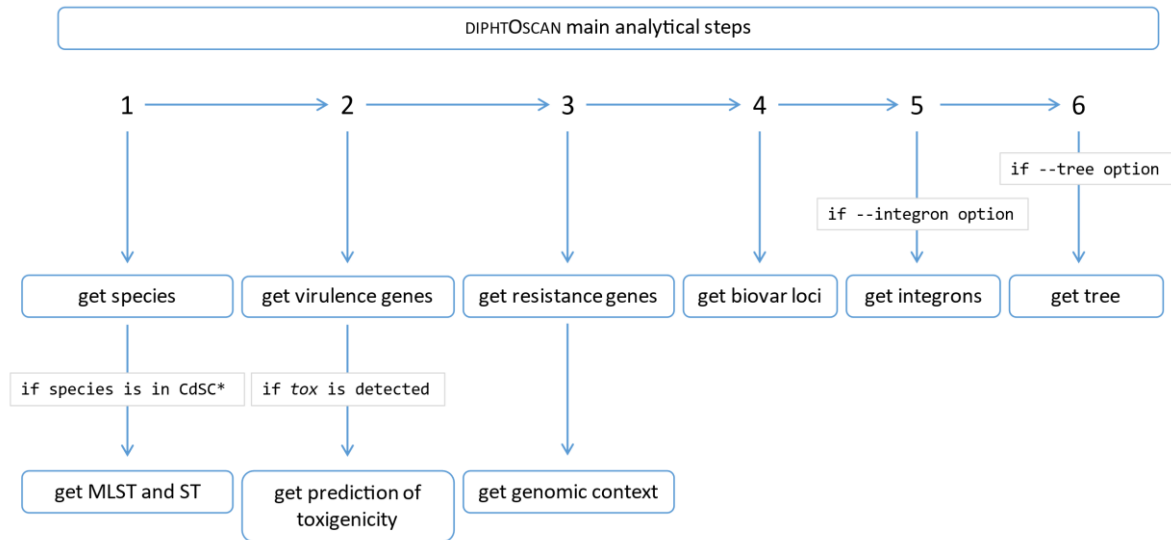
CdSC: Corynebacterium diphtheriae species complex

81

Figure S2. Flowchart of the main analytical steps of DIPHTOSCAN

DIPHTOSCAN starts with a taxonomic check, and then proceeds to virulence genes detection (including the tox gene, and its possible truncations), resistance features (genes and mutations, and their co-localization), and biovar-associated genes (*spuA* and *nar* clusters). Optionally, the presence of integrons is detected, and a k-mer distance-based tree can be built.
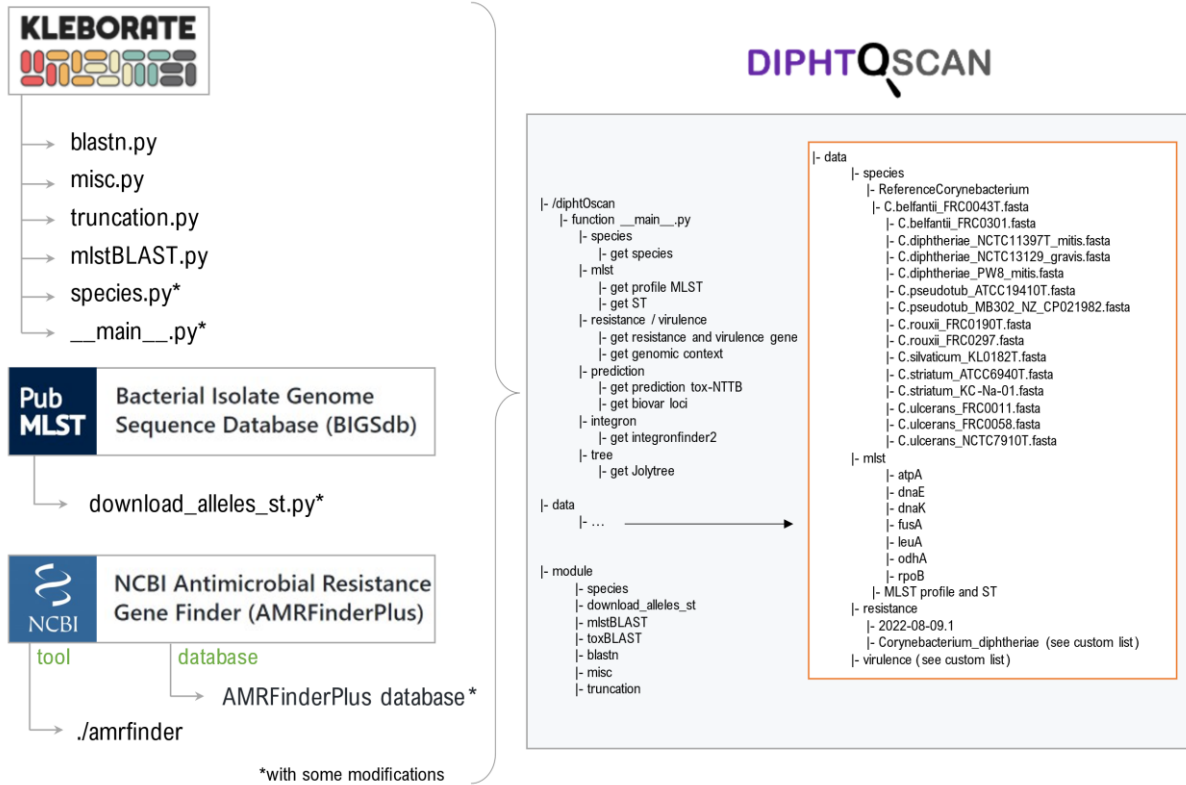
Figure S3. Bioinformatics tools integrated into DIPHTOSCAN

DIPHTOSCAN is a pipeline which groups a set of novel scripts and adaptations of previous tools, including Kleborate, BIGSdb, AMRFinderPlus and IntegronFinder2.

Custom features of DIPHTOSCAN

|- /AMRFinderPlus
  |- database
    |- \<database format version> e.g., 3.10
      |- Corynebacterium_diphtheriae
        |- fam.tab
        |- AMRProt
        |- AMRProt-mutation.tab
        |- AMRProt-suppress
        |- AMRProt-susceptible.tab
        |- AMR.LIB
        |- AMR_CDS
        |- AMR_DNA-\<organism>
        |- AMR_DNA-\<organism>.tab
        |- changes.txt
        |- min_software_version.txt
        |- version.txt
        |- changes.txt
        |- taxgroup.tab
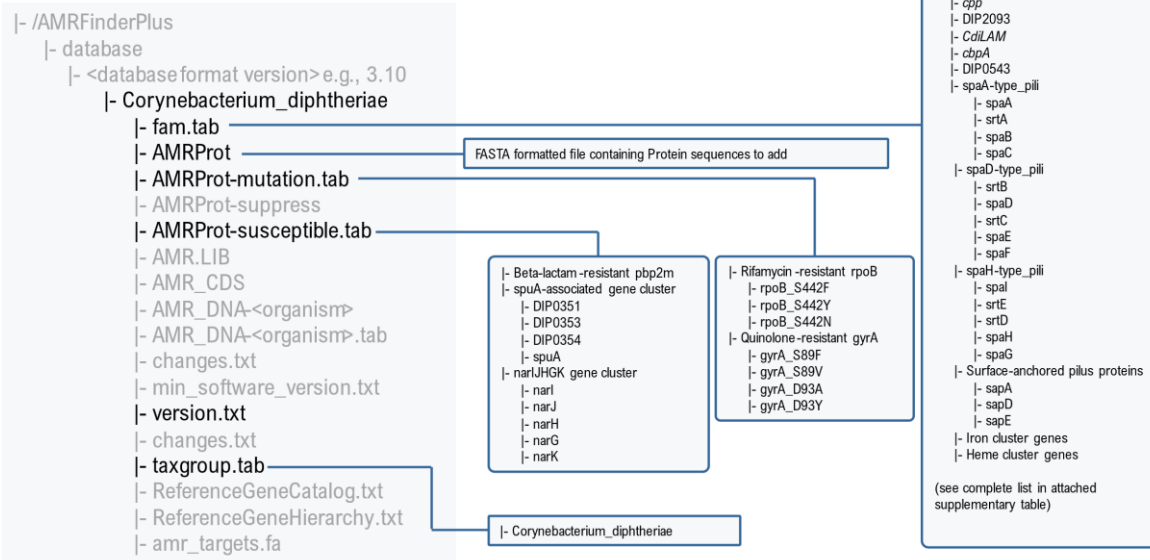 |- ReferenceGeneCatalog.txt
        |- ReferenceGeneHierarchy.txt
        |- amr_targets.fa

FASTA formatted file containing Protein sequences to add

|- Beta-lactam -resistant pbp2m
|- spuA-associated gene cluster
  |- DIP0351
  |- DIP0353
  |- DIP0354
  |- spuA
|- narIJHGK gene cluster
  |- narI
  |- narJ
  |- narH
  |- narG
  |- narK

|- Rifamycin -resistant rpoB
  |- rpoB_S442F
  |- rpoB_S442Y
  |- rpoB_S442N
|- Quinolone -resistant gyrA
  |- gyrA_S89F
  |- gyrA_S89V
  |- gyrA_D93A
  |- gyrA_D93Y

|- Corynebacterium_diphtheriae

|- VIRULENCE
  |- *tox*
  |- *dtxR*
  |- *pld*
  |- *rbp*
  |- *rbp -like*
|- ADHERENCE
  |- DIP0733
  |- DIP1281
  |- DIP1621
  |- DIP1880
  |- *RhuM*
  |- *cpp*
  |- DIP2093
  |- *CdiLAM*
  |- *cbpA*
  |- DIP0543
  |- spaA-type_pili
    |- spaA
    |- srtA
    |- spaB
    |- spaC
  |- spaD-type_pili
    |- srtB
    |- spaD
    |- srtC
    |- spaE
    |- spaF
  |- spaH-type_pili
    |- spaI
    |- srtE
    |- srtD
    |- spaH
    |- spaG
  |- Surface-anchored pilus proteins
    |- sapA
    |- sapD
    |- sapE
  |- Iron cluster genes
  |- Heme cluster genes

(see complete list in attached supplementary table)

**Figure S4. Custom features of DIPHTOSCAN database**

Description of the custom features of the DIPHTOSCAN database. The database was built by adding to the AMRFinderPlus database, *ad-hoc* resistance and virulence genes, as well as molecular determinants to predict the biovar.
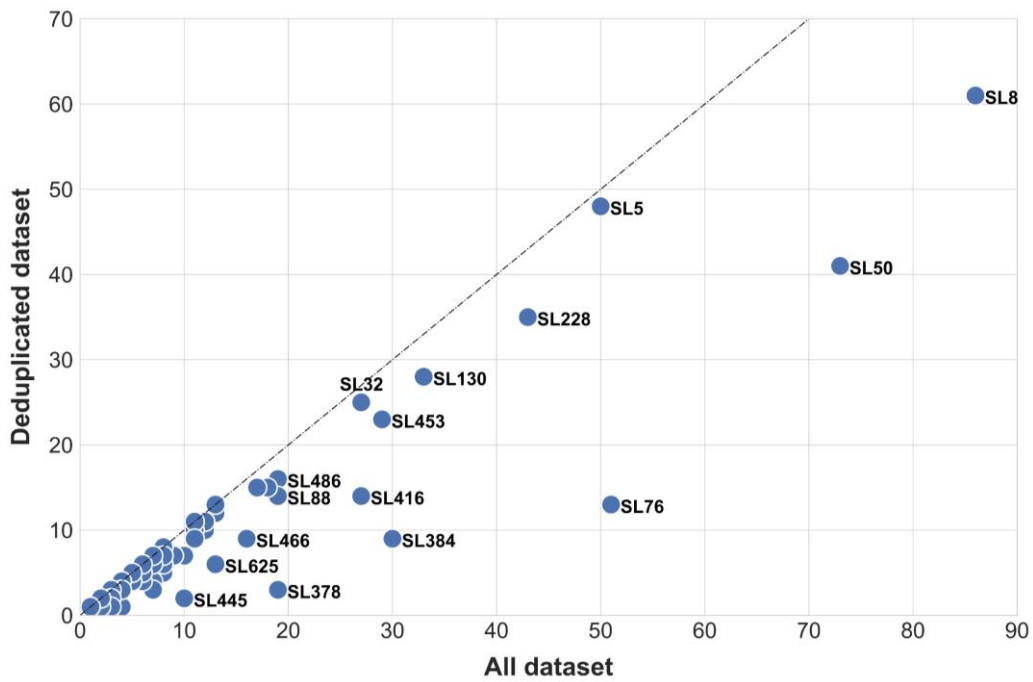
A. Effect of deduplication of the global dataset

B. Sublineage representation in France-2022 versus the deduplicated dataset.
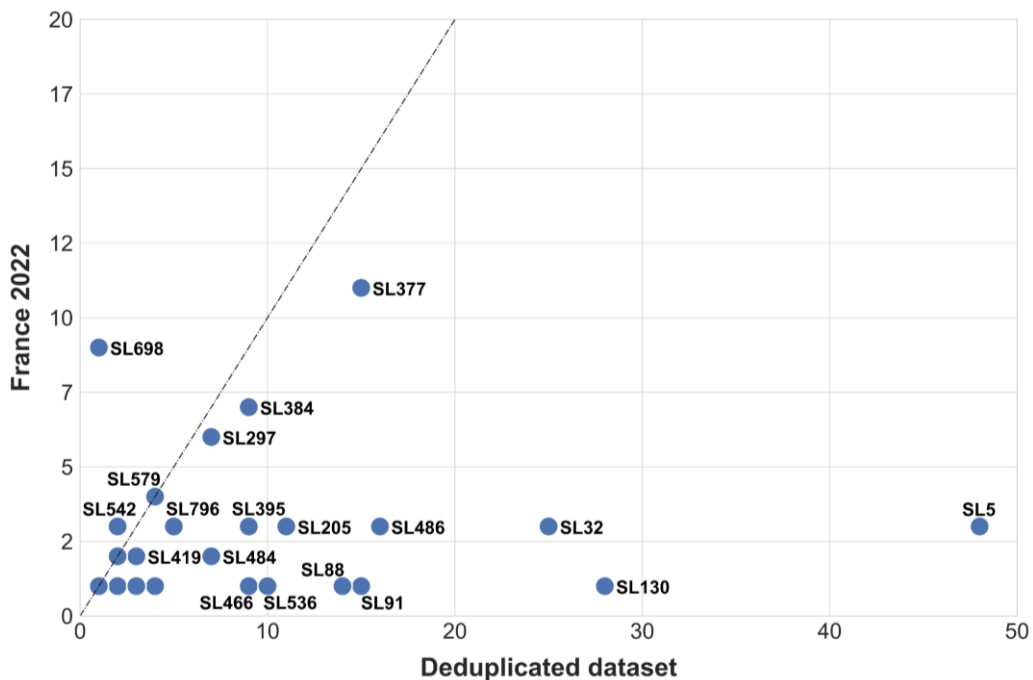
Figure S5. Sublineage frequency comparisons between datasets

A. Effect of deduplication of the global dataset. X-axis: The number of isolates in the entire global dataset (n=1,249 genomes). Y-axis: The number of isolates in the deduplicated global

150   dataset (n=976). Each sublineage is represented as a circle; the main sublineages' identifiers

151   are shown.

152   **B**. Sublineage representation in France-2022 *versus* the global (deduplicated) dataset. X-axis:

153   The number of isolates in the deduplicated dataset (n=976 genomes). Y-axis: The number of

154   isolates in the France-2022 dataset (n=101). Each sublineage is represented as a circle; the main

155   sublineages' identifiers are shown.

156

157

158

159

160

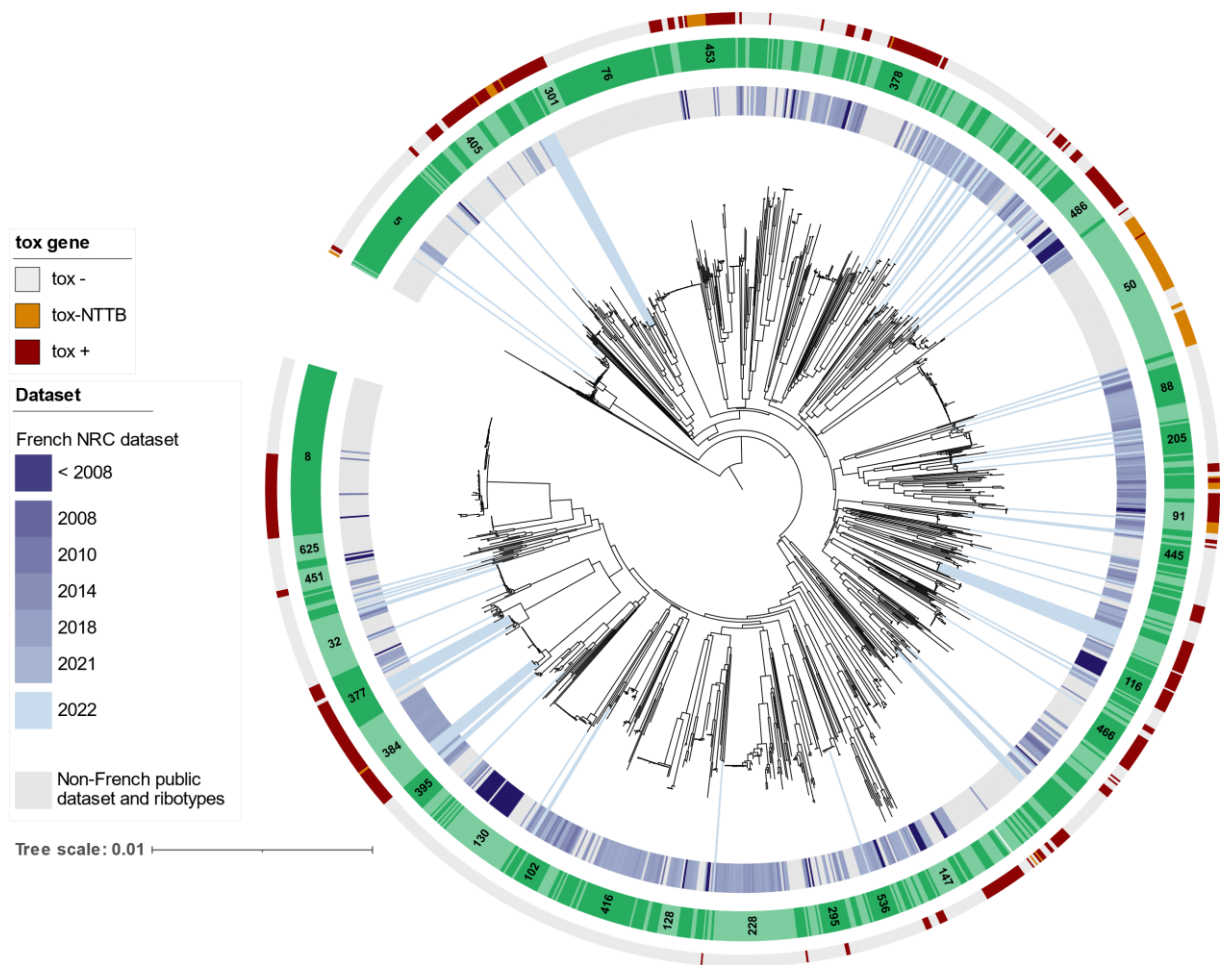161

162

163

164

165

166

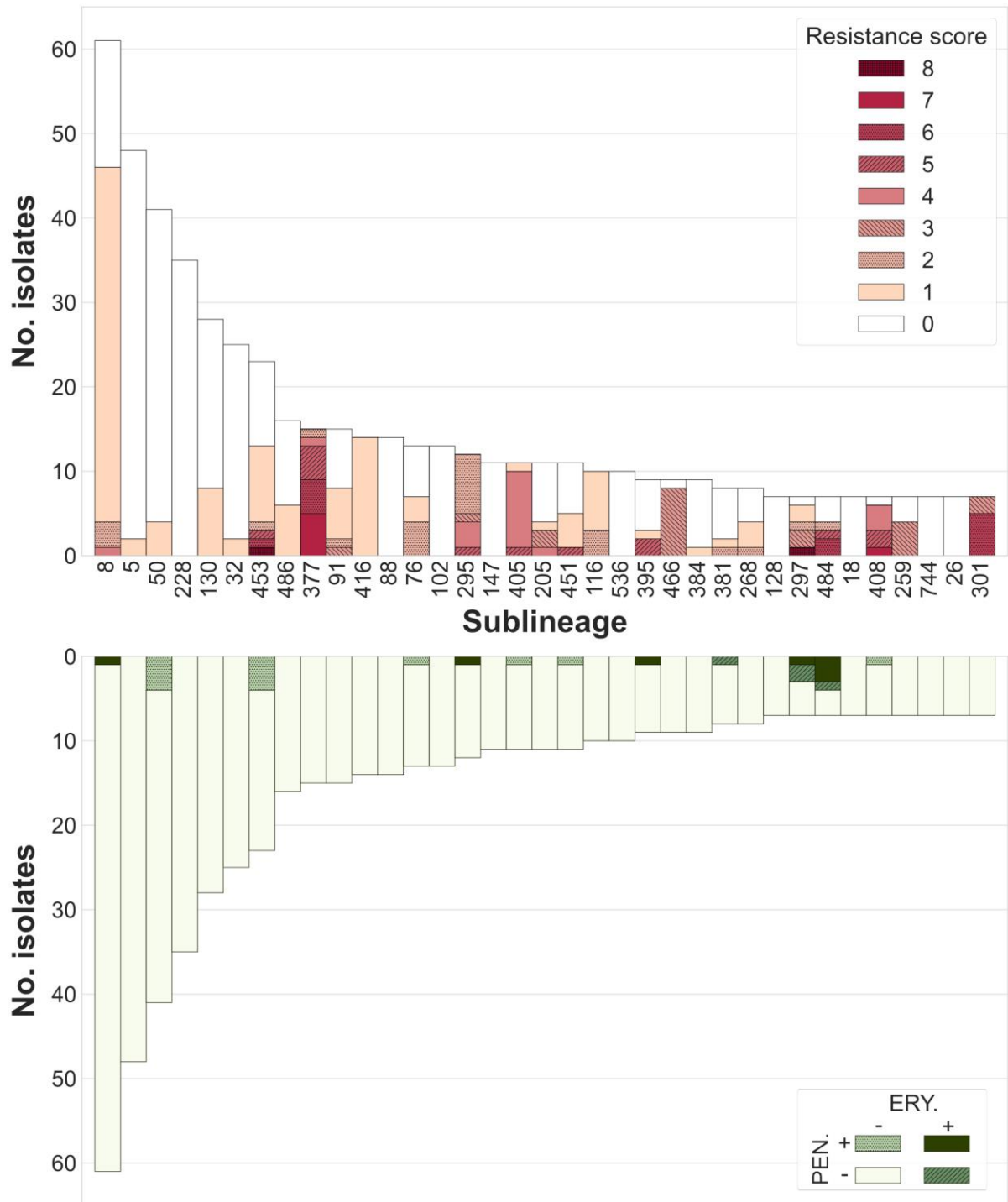167

168

169

170

171

172

173

174

Figure S6. **Phylogenetic tree of 1,350 *Corynebacterium diphtheriae***

The tree shows the placement of the France-2022 isolates (n=101) within the diversity of the 1,249 global non-deduplicated genomes. The tree was generated using JolyTree (Criscuolo, 2020). The scale bar represents the number of nucleotide substitutions per site. The first circle indicates the dataset, with France-2022 isolates indicated with light-blue rays from the tree leaves, and a blue color gradient (darker blue for older isolates) for French NRC isolates prior to 2022. Grey color sectors correspond to the global genomes. The second inner circle indicates sublineage alternation; main sublineages are labeled within the sectors. The following circle indicates the presence, disruption or absence of the diphtheria toxin *tox* gene (see key).

190

191

192 **Figure S7**. **Phylogenetic position of the ribotype reference strains of *Corynebacterium***

193 ***diphtheriae***

194 The tree was obtained by maximum likelihood based on a multiple sequence alignment of the

195 core genome, and was rooted with *C. belfantii* (not shown; same tree as in main Figure). The

196 scale bar gives the number of nucleotide substitutions per site. The inner circle indicates

197 sublineage alternation; main sublineages are labeled within the sectors. The second circle

198 indicates the presence, disruption or absence of the diphtheria toxin *tox* gene (see key). Each

199 point on the tree leaves corresponds to a ribotype reference strain. The ribotype names and

200 reference strain identifier are indicated outside the circles.

201

202

203

204

Figure S8. Sublineage distribution of resistance features, including *pbp2m* and *ermX*

(Top) The bars height corresponds to the number of genomes per sublineage (entire dataset, 1,249 isolates). Bars are colored according to the resistance score, i.e. the number de resistance families present in the strains (see color key).

(Bottom) Same as in A, with genomes carrying genes associated with resistance to penicillin (PEN.) and/or erythromycin (ERY.) resistance being colored (see key).

Tree scale: 0.01
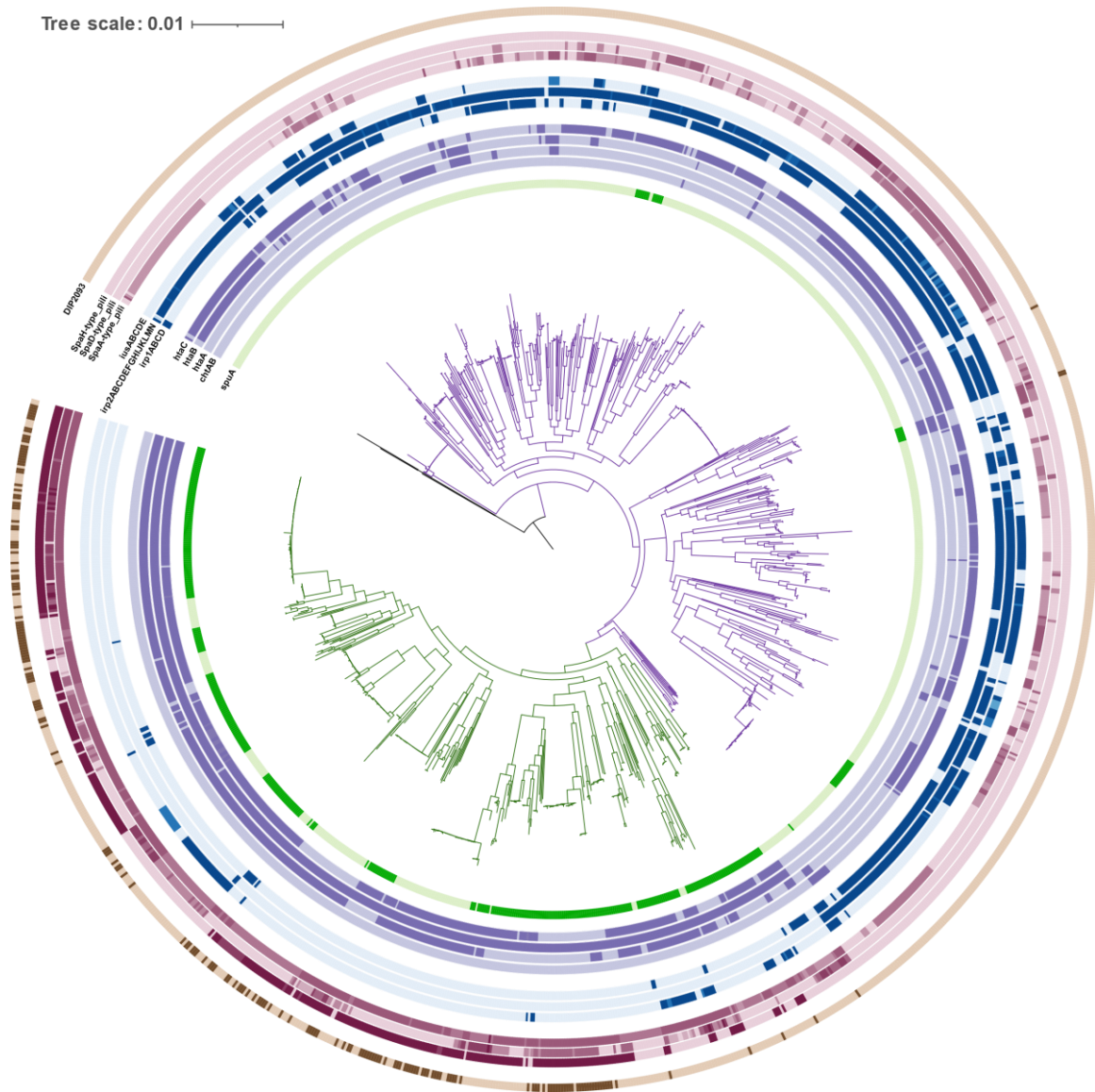
Figure S9. **Phylogenetic distribution of iron metabolism and adhesion-associated genes in**
Corynebacterium diphtheriae **lineages Mitis and Gravis**

The tree was obtained by maximum likelihood based on a multiple sequence alignment of the
core genome (as in main Figure). 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 first inner circle indicates the presence of the *spuA* gene (darker color:
presence). The four following circles indicate the presence of the heme metabolism-associated
genes *chtAB*, *htaA*, *htaB* and *htaC*. The three following circles indicate the presence of the iron
metabolism-associated genes: *irp1ABCD*, *irp2ABCDEFI-irp2JKLMN* and *iusABCDE*. Next,
three following circles indicate the presence of the fimbriae gene clusters: SpaA-type, SpaD-
type and SpaH-type. The most external circle indicates the presence of collagen-binding protein
DIP2093. For circles that represent clusters of genes, a color gradient represents the number of
genes present in the cluster (darker: more genes).