Genomic Epidemiology and Strain Taxonomy of Corynebacterium diphtheriae

Julien Guglielmini, a Melanie Hennart, b,c Edgar Badell, b,c Julie Toubiana, b,c,d Alexis Criscuolo, a © Sylvain Brisse b,c

a Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, Department of Computational Biology, Paris, France
b Institut Pasteur, Université de Paris, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France
c National Reference Center for the Corynebacteria of the Diphtheriae Complex, Paris, France
d Université de Paris, Service de Pédiatrie Générale et Maladies Infectieuses, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris, Paris, France
e Sorbonne Université, Collège Doctoral, Paris, France

ABSTRACT Corynebacterium diphtheriae is highly transmissible and can cause large diphtheria outbreaks where vaccination coverage is insufficient. Sporadic cases or small clusters are observed in high-vaccination settings. The phylogeography and short timescale evolution of C. diphtheriae are not well understood, in part due to a lack of harmonized analytical approaches of genomic surveillance and strain tracking. We combined 1,305 genes with highly reproducible allele calls into a core genome multilocus sequence typing (cgMLST) scheme. We analyzed cgMLST gene diversity among 602 isolates from sporadic clinical cases, small clusters, or large outbreaks. We defined sublineages based on the phylogenetic structure within C. diphtheriae and strains based on the highest number of cgMLST mismatches within documented outbreaks. We performed time-scaled phylogenetic analyses of major sublineages. The cgMLST scheme showed high allele call rate in C. diphtheriae and the closely related species C. bel扶antii and C. rouxi. We demonstrate its utility to delineate epidemiological case clusters and outbreaks using a 25 mismatches threshold and reveal a number of cryptic transmission chains, most of which are geographically restricted to one or a few adjacent countries. Subcultures of the vaccine strain PW8 differed by up to 20 cgMLST mismatches. Phylogenetic analyses revealed a short-timescale evolutionary gain or loss of the diphtheria toxin and biovar-associated genes. We devised a genomic taxonomy of strains and deeper sublineages (defined using a 500-cgMLST-mismatch threshold), currently comprising 151 sublineages, only a few of which are geographically widespread based on current sampling. The cgMLST genotyping tool and nomenclature was made publicly accessible (https://bigsdb.pasteur.fr/diphtheria). Standardized genome-scale strain genotyping will help tracing transmission and geographic spread of C. diphtheriae. The unified genomic taxonomy of C. diphtheriae strains provides a common language for studies of ecology, evolution, and virulence heterogeneity among C. diphtheriae sublineages.

KEYWORDS cgMLST, diphtheria, epidemiology, genomic epidemiology, genomics, microevolution

Corynebacterium diphtheriae causes diphtheria, a severe infection of the upper respiratory tract. The most typical symptoms include the formation of pseudomembranes in the throat and toxin-induced complications such as myocarditis and neuropathy. The diphtheria toxin is produced by some strains that carry the tox gene, harbored by a lysogenized corynephage. Cutaneous infections by C. diphtheriae are common, and skin carriage or infection may act as source of respiratory infections (1). Isolates without the tox gene can cause invasive infections such as endocarditis and sepsis (2). C. diphtheriae is considered to spread predominantly through droplets among humans and is only rarely isolated from animals. Strains of the phylogenetically related species C. ulcerans and C. pseudotuberculosis can also produce the
Diphtheria is largely eliminated in countries with high coverage rates of vaccination with the diphtheria toxoid (3). Nevertheless, the pathogen C. diphtheriae continues to circulate among humans and even in high-vaccination coverage countries, sporadic cases of toxigenic C. diphtheriae infections are reported regularly. In Europe, 139 cases were reported between 2014 and 2018, a majority of which followed introduction of C. diphtheriae from regions of endemicity (4). In these settings, transmission is rare, and only small clusters of patients or carriers are observed (5, 6). In contrast, in settings where public health action is disrupted and vaccination coverage has decreased, several large outbreaks of diphtheria in its typical clinical form have occurred in recent years (7–9).

Strain subtyping is used to study genotype-phenotype relationships and to define chains of transmission and the origin of outbreak strains. The subdivision of C. diphtheriae into four biovars is of poor resolution and has no epidemiological usefulness, and even its relevance for strain classification is questionable (10, 11). The recently described species C. belfanti (12) and C. rouxii (13) include most strains that were previously assigned to C. diphtheriae biovar Belfanti. Although no toxigenic strains of these two novel species have been described thus far, they are typically misidentified as C. diphtheriae when using reference molecular tests, biochemical diagnostic methods, and even matrix-assisted laser desorption ionization–time of flight mass spectrometry (12–14).

The highly standardized molecular methods ribotyping (15) and seven-gene MLST (14) have been developed to subtype C. diphtheriae, but in the genomic era their resolution power appears to be highly limited (11). Whole-genome sequencing (WGS) provides nearly ultimate resolutive power, and several recent studies have used this approach for C. diphtheriae population biology (10, 16–18) or molecular epidemiology and outbreak investigation (5, 19–23). However, currently a shared genotyping system and strain definition standard are lacking, which limits our ability to compare strains across studies.

Among WGS analytical approaches, core genome multilocus sequence typing (cgMLST) (24) has several attractive characteristics, including high reproducibility and portability that allow strain matching across studies, geography, and time. cgMLST has been used successfully to define transmission of C. diphtheriae among patients (5, 20, 21). However, at present, no cgMLST scheme is provided through a publicly available web application.

The aim of this work was to address this gap by providing an accessible genotyping standard and harmonized definitions of strains and sublineages. We calibrated the discriminatory power of cgMLST by comparing this proposed standard to whole-genome single nucleotide polymorphisms (SNPs) and illustrate how cgMLST can be used to identify strains from recent C. diphtheriae clusters and outbreaks. Last, we mapped seven-gene MLST sequence type (ST) identifiers to the novel cgMLST-based sublineage nomenclature, which provides backward compatibility with this previous classification framework.

MATERIALS AND METHODS

Isolates and outbreak sets included in the study. A total of 440 isolates (see Table S1 in the supplemental material) were included into the study data set. First, we included and sequenced eight isolates of the French National Reference Center (NRC) collection corresponding to four available pairs of C. diphtheriae infection cases defined as epidemiologically linked (Table 1). The first pair (Mayotte set 1) included two isolates from a single sample, isolated 1 month apart (FRC025 and FRC030). The second pair (Mayotte set 2) included isolates from one patient and one of his contacts; the two isolates were isolated 15 days apart. The third pair (Carcassonne set 1) included isolates from two members of the same household, sampled 14 days apart. The fourth pair (Boulogne set 1) included two isolates from a pair of siblings, sampled on the same day. Second, we included and sequenced a random selection of 21 isolates from the French NRC collection, selected among isolates that were cocirculating temporally or geographically with each outbreak (see Table S1, “sporadic” with “FRCxxxx” isolate names). Third, 172 isolates from previously published genomic epidemiology studies in South Africa (25), Germany-Switzerland (5), Germany (20), France (22), Australia (17), and Canada (21) were included. These genomic investigations of outbreaks (or clusters of cases) together comprised 11 sets of isolates defined as related based on epidemiologically and genomic evidence (Table 1) and also included isolates considered nonrelated to the outbreak sets. Finally, all other publicly available genomic sequences from GenBank available 30 May 2019 were included (see Table S1). Of these, we removed isolates from the Canadian study (isolates CD46 to CD49 and isolates CD51 to CD56) that were of low quality. We also removed three isolates with atypical assemblies: two from Germany (KL0932 and
| Data set                  | No. of strains | 7-Gene MLST | Sublineage | Genetic cluster | Country          | Yr(s)          | Outbreak information                                                                 | Source or reference |
|--------------------------|----------------|-------------|------------|-----------------|------------------|----------------|--------------------------------------------------------------------------------------|---------------------|
| South Africa set 1       | 19             | 378         | 378        | 472             | South Africa     | 2015–2016     | tox positive                                                                                   | 25                  |
| South Africa ST395       | 5              | 395         | 395        | 68;491;492;542  | South Africa     | 1980; 2015    | tox negative                                                                                   | 25                  |
| Germany-Switzerland set 1 | 3              | 473         | 473        | 13              | Germany          | 2015           | Refugees from Eritrea, Ethiopia, Somalia, and Syria; tox positive                           | 5                   |
| Germany-Switzerland set 2 | 3              | 183;486     | 486        | 204             | Germany          | 2015           | Refugees from Eritrea, Ethiopia, Somalia, and Syria; tox positive                           | 5                   |
| Germany-Switzerland set 3 | 4              | 398         | 398        | 207             | Germany          | 2015           | Refugees from Eritrea, Ethiopia, Somalia, and Syria; tox positive                           | 5                   |
| Germany cluster 1        | 14             | 8           | 8          | 414             | Germany          | 2016–2017      | Hamburg, Essen, Kiel                                                                         | 20                  |
| Germany cluster 2        | 19             | 8           | 8          | 292             | Germany          | 2015–2017      | Berlin mostly                                                                                   | 20                  |
| Germany cluster 3        | 6              | 8           | 8          | 411             | Germany          | 2015–2016      | Berlin, Essen, Trier                                                                         | 20                  |
| Germany cluster 4        | 2              | 8           | 8          | 415             | Germany          | 2016–2017      | Hannover                                                                                       | 20                  |
| Germany ST130            | 10             | 130         | 130        | 402             | Germany          | 2012–2017      | None                                                                                           | 20                  |
| Vancouver cluster 1      | 43             | 76;746      | 76         | 456             | Canada           | 2017–2018      | Vancouver inner city, cutaneous infection cluster                                          | 21                  |
| Mayotte set 1            | 3              | 97          | 97         | 78              | Mayotte          | 2008–2009      | Single sample, 1 mo apart                                                                     | This study          |
| Mayotte set 2            | 2              | 184         | 184        | 86              | Mayotte          | 2009           | Patient + contact, 15 days apart                                                                | This study          |
| Carcassonne set 1        | 2              | 212         | 212        | 222             | France           | 2011           | Members of household, 14 days apart                                                            | This study          |
| Bron set 1               | 2              | 411         | 411        | 116             | France           | 2016           | Siblings, sampled on same day                                                                  | This study          |
KL0954) because they were highly fragmented (446 and 496 contigs, respectively) and one from a Germany-
Switzerland study (Cd_14) because of a low G+C content (42.9%).

While this work was in progress, additional C. diphtheriae genomes were published from Austria (26),
the United States (23), Yemen (9), France (11), Spain (27), and India (28). We included these 162 genomes
in a joint phylogenetic analysis with the 440 genomes indicated above (see Fig. S1).

DNA preparation and sequencing. Isolates were retrieved from −80°C storage and plated on tryp-
tose-casein soy agar for 24 to 48 h. A small amount of bacterial colony biomass was resuspended in a
lysis solution (20 mM Tris-HCl [pH 8], 2 mM EDTA, 1.2% Triton X-100, and lysozyme [20 mg/ml]) and incu-
bated at 37°C for 1 h, and DNA was extracted with the DNeasy Blood&Tissue kit (Qiagen, Courtaboeuf,
France) according to the manufacturer’s instructions. Genomic sequencing was performed using a
NextSeq500 instrument (Illumina, San Diego, CA) with a 2 × 150-nucleotide (nt) paired-end protocol fol-
lowing Nextera XT library preparation (11).

Genomic sequence assembly. For de novo assembly, paired-end reads were clipped and trimmed using
AllenTrimmer v0.4.0 (29), corrected using Musket v1.1 (30), and merged (if needed) using FLASH v1.2.11
(31). For each sample, the remaining processed reads were assembled and scaffolded using SPAdes v3.12.0 (32).

Definition of the cgMLST scheme. We first selected 171 genomes representing a diversity of C. diphtheriae
sublineages, including the 104 genomes presented in Table S1 and 67 genomes from the French NRC (11).
These included 14 C. belfanti isolates, but no C. rouxi isolates were included at this step, since this species had
not been not identified at that time. From this set of genomes, we inferred the species core genome using
the CoreGeneBuilder pipeline (https://zenodo.org/record/165206#.WVp7T95SEQ; see reference 33) using
the NCCT13129 strain genome (GenBank accession no. NC_002935; see reference 34). CoreGeneBuilder automatically
removed genomes for which the size was too divergent compared to the whole set. We applied the following
quality filters: a maximum of 500 contigs and a minimum N50 of 10,000 bp. With these criteria, no genome was
filtered out. The pipeline’s next step relies on eCAMBer (33), which consists of a de novo annotation of the
genomes (except the reference) using Prodigal (36) and in the harmonization of the positions of the stop and
start codons. In the last step, the core genome is inferred using a bidirectional best hits approach, following
Touchon et al. (37). We used CoreGeneBuilder default settings and considered a gene as part of the core genome
if it was found in at least 95% of the 171 selected genomes. This resulted in an initial core genome containing
1,555 loci.

To obtain a set of loci that would be robust to genotyping artifacts, we filtered out some genes based on several criteria. First, we removed potential paralogs, since the presence of paralogs inside a typing scheme can lead to genotyping errors when a gene sequence is attributed to the wrong locus. To detect the potential paralogs, we compared each allele of each locus against all the alleles of all the other loci using the software BLAT (38). If a single hit was found between two different loci (more than 70% amino acid sequence identity between two alleles), we removed both. A total of eight loci were discarded by this process. We also removed genes that belong to the classical seven-gene MLST scheme (14) and the ribosomal protein genes that are used in the ribosomal MLST approach (39), so that they could be analyzed independently. Next, we removed loci whose length varied too much among alleles, which is useful to reduce ambiguities during the genotyping process. We aligned the amino acid sequences and removed loci for which the alignment contained more than 5% of gaps (total number of gaps compared to the total number of character states). This filter removed 107 loci. In addition, we removed three loci because at least one allele showed one or more ambiguous character state(s). Altogether, these filters resulted in a list of 1,388 remaining loci.

To filter out loci with nonreproducible allele calls at low assembly coverage depth, randomly selected read
pairs representing coverage depths of 10× to 50× were drawn (five times for each isolate/coverage combina-
tion) from the raw sequencing data for four isolates (C. diphtheriae 18-093a and FRC0024 and C. belfanti FRC0019
and FRC0318); we included two C. belfanti isolates because their genome is more fragmented due to a high
copy number of insertion sequences (40). After assembly and allele calling using BIGSdb, we eliminated 33 loci
that showed nonreproducible allele calls from all assemblies obtained with coverage depths equal to or greater
than 20× in any of the four isolates.

Finally, we also removed 52 loci that were uncalled in >10% of scanned genomes. These filtering
steps led to a final set of 1,305 gene loci, which together constitute the “Pasteur C. diphtheriae cgMLST
scheme” (abbreviates here as “cgMLST scheme”). Note that a given gene locus could have been filtered out by
several filters applied independently, so that the sum of numbers of loci removed by individual filters is smaller than
the final number of loci that were filtered out.

Evaluation of cgMLST allele call rate. Because loci with high numbers of missing alleles were filtered
out during cgMLST scheme construction, the allele call rate of the resulting scheme was assessed using
an independent set of genomes, which had not been used for the above-described definition of
the scheme. This entirely distinct data set of C. diphtheriae isolates was made of 162 genomes published after
cgMLST scheme definition and described in the first Materials and Methods paragraph. Missing alleles
were defined as being not called using 90% identity and 90% length coverage.

Constitution of a BIGSdb database for C. diphtheriae. The cgMLST scheme was incorporated into a
BIGSdb seqdef database (41). Classification schemes were created at the level of 25 and 500 mis-
matches. The isolates provenance data and genome sequences were imported into a linked isolates database
(https://bigdb.pasteur.fr/diphtheria). A publicly available BIGSdb project (i.e., a browsable list of
isolates) entitled “cgMLST set-up study” corresponds to the data set of 440 genomes and was created to
facilitate retrieval of the data used in this work.

Inheritance of classification identifers from seven-gene MLST. MLST alleles and profiles of the
440 genomes are defined in the PubMLST C. diphtheriae database. We then mapped the sequence type (ST)
identifiers onto the obtained 500-mismatch cgMLST single-linkage classification groups and attributed the
dominant ST identifier to the group, using an inheritance algorithm (https://gitlab.pasteur.fr/BEBP/Inheritance-algorithm).

Comparison to other schemes. An ad-hoc cgMLST scheme for C. diphtheriae was defined and used previously (19, 20, 42). In order to compare our scheme with this previous one, we took the first allele of all loci of both schemes and performed an all-against-all BLASTN analysis. Hits were defined as sequence pairs with a proportion of identical nucleotides greater than 90% and with an alignment length longer than half of both the query and the subject sequences.

Cryptic cluster detection. Cryptic clusters were defined as single-linkage groups of genomes (i) defined using a threshold of 25 cgMLST allelic mismatches but (ii) which were not initially defined as clusters based on available epidemiological data.

Phylogenetic and population structure analyses. Phylogenetic analyses were performed by first inferring a tree using JolyTree v1, an alignment-free distance-based procedure (43). The resulting tree was used as input for ClonalFrameML (44), which also requires an alignment. The alignment was obtained by translating allelic sequences of cgMLST loci of each genome into proteins. We aligned the proteins of each loci using MAFFT v7.467 (45) with default parameters and then back-translated the protein alignments to codon alignments using Goalign (https://github.com/evolbioinfo/goalign); the individual gene alignments were then concatenated into a supermatrix, also using Goalign (the nucleotide positions of missing alleles were replaced by gaps in the concatenate of alignments of individual loci. We used this alignment and above defined input tree to run ClonalFrameML with default parameters. We used the software PHIPack v1.0 (46) to calculate the PHI (pairwise homoplasy index) statistic and perform a locus-by-locus test of recombination with a P value threshold of 0.05. Nucleotide diversity was calculated both for the entire data set of each species and among sublineages using an in-house script relying on the \( d \)-distance calculation provided by Goalign. The overall average Silhouette width, which estimates the overall clustering quality based on maximizing intergroup distances and minimizing intra-group distances, was used to define the clustering consistency (47).

Dating ST5 and ST8 phylogenetic tree nodes was performed using the LSD (least-squares dating) algorithm (48) as implemented in IQ-TREE v2.0.6 (49). The sampling dates were used to calibrate the trees. The subtrees were extracted from the entire tree (Fig. 1). For both analyses, the closest outgroup was also extracted and included in the analysis. This allowed for the dating of the root of the ST8 phylogeny, whereas for ST5 this analysis did not allow dating the root.

The clade compatibility index of STs or other groups was calculated using the ETE Python library (http://etetoolkit.org/docs/latest/tutorial/tutorial_trees.html#checking-the-monophyly-of-attributes-within-a-tree) in order to define whether their constituent genomes formed a monophyletic, paraphyletic or polyphyletic group within the recombination-purged sequence-based phylogeny of the core genome. We estimated clade compatibility as the proportion of nonsingleton STs or sublineages that were monophyletic.

Whole-genome SNP calling based on a read mapping approach. We used Snippy v4.4.0 (50) with default parameters to perform SNP calling within the German and Canadian outbreak data sets, using strains NCTC13129 and CD50 (from BioProject PRJNA563223), respectively, as references.

Saturation analysis of the number of sublineages and genetic clusters. Following the definitions of genetic clusters and sublineages (25 allelic mismatches for the genetic clusters and 500 allelic mismatches for the sublineages), we simulated 1,000 samples of 1 to 440 profiles (the 440 profiles of our initial data set), with a step of 1. For each threshold and each number of profiles, we then calculated the mean number of clusters and used the minimum and maximum values to define the range. To extrapolate the previously obtained curves, we defined the species accumulation curves using the R package vegan. We chose the Lomolino function (51) for its excellent fit with our data.

Data availability. Newly sequenced isolates have been deposited in the European Nucleotide Archive (accession numbers ER56637895 to ER56637897).

RESULTS

Definition and evaluation of a cgMLST scheme for genotyping isolates of C. diphtheriae, C. belfanti, and C. rouxi. We defined a set of 1,305 protein-coding gene loci deemed appropriate for C. diphtheriae, C. belfanti, and C. rouxi genotyping (see Materials and Methods). These loci were combined into a cgMLST scheme and altogether have a total length of 1,265,478 nt (calculated for allele 1 of each gene), representing 51% of the genome size of the ST8 reference strain NCTC13129. Individual locus lengths varied from 93 to 4,008 nt. The list of loci and their characteristics are presented in Table S2 in the supplemental material.

The allele call rate of the cgMLST scheme loci was determined using posteriorly published genomes from Austria (26), the United States (23), Yemen (9), France (11), Spain (27), and India (28) and was found to be 98.6% ± 1.5%. We also calculated the call rate using 337 additional, prospectively collected genome sequences from the French NRC for diphtheria (unpublished). The mean numbers of called alleles were 1,288 (98.7%) and 1,287 (98.6%) for C. diphtheriae and C. belfanti, respectively. We noted that a few C. belfanti genomes had a lower call rate (see Fig. S2). For C. rouxi, the allele call rate was lower (1,190; 91.2%), as expected, since this novel species was not included during scheme development and is
FIG 1  Phylogenetic tree of *C. diphtheriae* sublineages. A maximum-likelihood phylogenetic tree for 440 *C. diphtheriae*, *C. belfanti*, and *C. rouxi* isolates was developed based on the concatenated multiple sequence alignments of 1,305 core genome multilocus sequence typing loci. The top panel represents the unrooted topology of the three species. The bottom panel is the *C. diphtheriae* subtree. Branches with support values greater than 70% are thicker. The root of this subtree corresponds to the placement of the branch leading to *C. rouxi* and *C. belfanti* in the top panel tree. The *C. diphtheriae* subtree (Continued on next page)
genetically more distant (13). The allele call rate was not significantly different between tox-positive and tox-negative *C. diphtheriae* (98.8 versus 98.6%, respectively; see Fig. S3). The relationships between allele call rate and genome assembly fragmentation showed that high call rates (>95%) were generally obtained, even for *C. diphtheriae* genome assemblies that were fragmented into >400 contigs (see Fig. S2). Overall, these results demonstrate high typeability across the phylogenetic breadth of *C. diphtheriae* and its two closely related species, *C. belfantii* and *C. rouxii*.

Based on the study data set of 440 genomes, allele numbers ranged from 7 to 285 per locus, illustrating the large amounts of genetic diversity of the core genes within *C. diphtheriae*. As expected, the number of alleles per locus increased with locus size (see Fig. S4). Locus-by-locus recombination analyses detected evidence of intragenic recombination within a majority of loci (843; 64.6%), whereas only 431 (33%) loci were deemed non recombinogenic. There were 31 (2.4%) loci with too few polymorphic sites to be tested (see Table S2). The number of alleles per locus was higher for recombinogenic loci (see Fig. S5), which reflects the contribution of homologous recombination to the generation of allelic diversity.

We compared the cgMLST scheme defined here (named the Pasteur scheme) to a previously published ad hoc (not publicly available for comparison) cgMLST scheme (20) and found 1,132 loci in common, representing 87% of the Pasteur cgMLST loci (see Fig. S6). Of these, 115 varied slightly in length due to alternative start codon definitions. Whereas 173 loci were unique to the Pasteur scheme, 421 were unique to the previous scheme. Three loci were not counted as common to the two schemes, even though their sequences matched, because they are encoded on opposite strands.

**Population structure of *C. diphtheriae***. A phylogenetic analysis was conducted based on a recombination-purged concatenate of the individual alignments of the 1,305 loci (Fig. 1). The three deepest branches corresponded to the three species *C. rouxii*, *C. belfantii*, and *C. diphtheriae*. Whereas *C. diphtheriae* included most isolates (86.5%, 148/171), *C. rouxii* and *C. belfantii* comprised 6 and 17 isolates, respectively. MLST data (see Fig. S7) revealed 134, 10, and 3 STs within *C. diphtheriae*, *C. belfantii*, and *C. rouxii*, respectively. The nucleotide diversity values, estimated as the average SNP distance among isolates, were 0.012 for *C. diphtheriae*, 0.0025 for *C. belfantii*, and 0.0039 for *C. rouxii*. The average genetic distance between sublineages (i.e., considering only one representative genome of each sublineage, as defined below) was 0.024 polymorphism per nucleotide position.

The phylogenetic structure within *C. diphtheriae* revealed a star-like phylogeny with multiple deeply branching sublineages as previously reported (11, 18). The phylogenetic distribution of seven-gene MLST data showed that sublineages typically corresponded to a single ST (Fig. 1). Some sublineages comprised closely related isolates, likely to be the result of sampling from case clusters, outbreaks or recent clonal expansions (see Table S1).

Diphtheria toxin tox gene-positive isolates were observed in 19 sublineages distributed on the Southern hemisphere of the tree presented in Fig. 1, including the sublineage that includes the vaccine strain PW8, and in five sublineages in the Northern hemisphere of the tree, including the one that comprised strain NCTC13129 (11). These two hemispheres of the tree are identifiable by the distribution of the *spuA* gene, involved in starch utilization (S2): the upper lineage was *spuA*-positive with the exception of six *spuA*-negative sublineages, whereas in sharp contrast, the lower lineage was almost entirely *spuA* negative, with ST5 being the single exceptional *spuA*-positive sublineage. The two major lineages of *C. diphtheriae* were therefore labeled as lineages Gravis and Mitis, respectively (11). The addition of 162 recently published genomes (see Fig. S1) contains the following annotations: the stars represent the presence (red star) or absence (white star) of the diphtheria toxin tox gene. The small circles represent the presence (green circle)/absence (white circle) of the *spuA* gene. The green strips represent the alternation of seven-gene MLST STs of the isolates: when the ST changes while walking along the circle, the shade of green also changes. Identifiers of the main STs are indicated. When an ST could not be attributed (because at least one of the seven genes was missing), the color is white. A strip with a visible border represents isolates with an ST that is different from the rest of the clade, and the ST number is provided below. For example, ST185 is found within the ST8 clade. The red strips follow the same alternation logic for the cgMLST sublineage identifiers. Note the strong concordance between ST and cgMLST sublineages. The position of noteworthy reference isolates is given by black arrows. Finally, the epidemic cluster positions are indicated by colored boxes on the corresponding branches (see bottom left key).
fully confirmed these observations. In addition to ST5, sublineage ST466 belonged to the lineage Mitis and included *spuA*-positive isolates. These *spuA*-positive isolates were closely related to two *spuA*-negative isolates, and all ST466 genomes were from India, suggesting a recent acquisition in this country of the *spuA* gene, along with probable conversion of the biovar from Mitis to Gravis.

**Definition of *C. diphtheriae* sublineages and their geographic distribution.** To guide the definition of *C. diphtheriae* sublineages, we performed an analysis of the mathematical properties of the clustering of cgMLST profiles at all possible thresholds (Silhouette analysis; see Fig. S8). We observed optimal clustering indices on a large interval (~between 300 and 600 mismatches), with a maximum value at 554. Since the Silhouette index was almost at maximum at 500 cgMLST mismatches, we chose this value for simplicity to define *C. diphtheriae* sublineages (SL). Our data set of 440 genomes comprised 114 sublineages, and a saturation curve analysis (Fig. 2) indicates that future addition of *C. diphtheriae* genomes is expected to reveal the existence of a multitude of additional sublineages that remain to be sampled. The additional data set of 162 genomes uncovered 22 additional sublineages (see Fig. S1). The average core genome nucleotide divergence among sublineages was 2.4%.

To create continuity in the numerical identifiers of sublineages with respect to MLST, we mapped the ST nomenclature on sublineages and incorporated the resulting sublineage tags into the BIGSdb *C. diphtheriae* genotype nomenclature database (see Table S3). Note that there was a strong concordance between MLST and cgMLST sublineages (see Fig. S9) and that this correspondence was established up to ST750 (17 November 2020).

All nonsingleton cgMLST sublineages (66/66; 100%) were compatible with the phylogeny, whereas this was the case for only 53 of 66 (80.3%) of seven-gene STs: 12 STs were polyphyletic, and 1 ST was paraphyletic.

The global distribution of *C. diphtheriae* sublineages, apparent from the current sample of 602 genomes, revealed a strong phylogeographic pattern with most sublineages being typically restricted to one or a few adjacent countries (Fig. 3). For example, SL8 was restricted...
to Europe, SL405 to India and SL40 to Indonesia. A notable exception was SL32, which was observed in Australia, Italy, and the UK. SL5 was isolated in Europe except one isolate from Australia.

**Frequently sampled sublineages of *C. diphtheriae***. The most frequent sublineages in our data set included seven-gene MLST genotypes ST5, ST8, ST76, and ST378; they were labeled as SL5, SL8, SL76, and SL378, respectively, by our inheritance script. These sublineages were previously reported from diphtheria outbreaks. SL378 comprised 19 tox-positive isolates (except for 1 tox-negative isolate) recovered during a 2015-2016 outbreak in the KwaZulu province of South Africa (25). Sublineage SL76 comprised isolates that were all tox-negative and belonged to ST76 or ST746 (a single-locus variant of ST76). Forty-five isolates were drawn from the Vancouver 2015-2018 outbreak of tox-negative *C. diphtheriae* (21), and five ST76 isolates were recovered from Belarus between 2004 and 2014 (see Table S1).

Sublineage SL5 comprised 39 tox-negative genomes belonging to ST5 (all except 6), ST728, and ST741 and 4 genomes for which the ST cannot be defined because they lack a *leuA* allele call (Fig. 4A). ST5 contributed to infections in Belarus from the mid-1990s until at least 2013 (16) and was also recovered from Italy, Australia (see Table S1), and Russia (14, 53). As noted above, whereas a majority of ST5 genomes were *spuA* negative, a distinct sub-branch within SL5 possessed the biovar Gravis marker *spuA*.

The phylogenetic analysis of SL5 genomes suggests initial acquisition of *spuA* in the branch leading to the SL5 ancestor (before approximately year 1834), followed by its loss in a ST5 sub-branch before 1955 (Fig. 1 and Fig. 4A).

SL8 was isolated from cases of the large ex-Soviet Union outbreak (34), and the genomes of this sublineage were isolated over a period of 20 years from Belarus, Germany, Russia, and the UK (see Table S1) (34, 28, 16, 20, 54). Strikingly, SL8 isolates were grouped into two sub-branches, which were either tox positive (sub-branch A) or tox negative (sub-branch B; Fig. 1 and Fig. 4B). Phylogenetic sister groups of the ST8 sublineage were ST123, ST451, and ST441, which comprised only tox-negative isolates. These data suggest a scenario of acquisition of the tox gene in the branch leading to the SL8 ancestor (before ~1967) and its subsequent loss in sub-branch B (before ~1988).
We aimed to assess the cgMLST variation within sets of isolates defined as epidemiologically related, i.e., described as case clusters or outbreaks (here, clusters) based on their provenance information. Fifteen such sets were collated, including four clusters identified by the French national surveillance system and sequenced for the purposes of this study (Table 1). The 15 clusters of infections were caused by diverse strains, which were largely scattered across the phylogeny of *C. diphtheriae*.

The distribution of allelic mismatches was clearly distinct for clusters versus nonrelated sets of isolates (Fig. 5), and whereas the number of allelic mismatches observed within clusters ranged from none to 24, allelic variation among nonrelated isolates typically had more than 1,000 mismatches.

To compare the discriminatory power of cgMLST with the expectedly higher resolution provided by whole-genome SNPs, two outbreaks for which read sets were available were investigated. The pairwise number of SNPs versus cgMLST allelic mismatches showed a strong correlation and indicated that 30 SNPs are observed between pairs of genomes that differ by 15 to 17 cgMLST mismatches (Fig. 6). Minimum spanning tree analysis of the genomes showed a high concordance between both methods in inferring genetic relationships of closely related genomes (see Fig. S10 and S11).

cgMLST cluster classification uncovers cryptic transmission and cryptic duplicated cultures. We classified the 440 *C. diphtheriae* genomes into single linkage groups using a threshold of 25 cgMLST mismatches, corresponding to the maximum value observed thus far for well-described clusters. We define these groups as genetic clusters. This analysis enabled to detect 43 cryptic clusters, i.e., isolates belonging to genetic clusters...
but for which we have no evidence of epidemiological links. The cryptic clusters were genetically more heterogeneous than the epidemiological clusters (Fig. 5). One cryptic cluster corresponded in fact to three distinct genome assemblies of the reference strain Park-Williams 8 (generally named PW8), which were deposited separately in public sequence databases (see Table S1). These genomes differed by up to 20 loci among them, suggesting
that multiple genetic differences exist among subcultures of this historical strain, isolated in 1896 and used for vaccine production. Likewise, there were four separately submitted genome sequences of strain C7beta (see Table S1), but these showed only $<$3 cgMLST distinct alleles. Cryptic cluster 8 included NCTC3529 and NCTC5011 (which are synonyms) and strain c21 from India, which differed from the two others by 22 loci.

Of the remaining 40 cryptic clusters, 33 were single-country clusters and may correspond to local chains of transmission that were previously undetected or just unreported to our knowledge. Only five cryptic clusters included isolates from at least two countries. Cluster 1 (as numbered in Table S1) corresponded to isolates from Belarus and the UK isolated between 1996 and 2004, and belong to the SL8 sublineage involved in the large ex-Soviet Union countries outbreak. Cryptic cluster 26 included two isolates from year 2008, one from Belarus and one from France, and the French patient had a history of travel to Russia; this cluster also belonged to SL8. Cryptic cluster 5 involved two isolates, one from Berlin in 2017 and one from Rio de Janeiro in 1999. Cryptic cluster 15 involved isolates from 2015 and 2016 in Australia and Germany. Cryptic 38 included isolates from France, Germany, and Australia (from 2013 to 2015). Cryptic cluster 3 involved two strains from Brazil, which were isolated in 1981 and 1993. Finally, we also revisited the clades defined in a recent study of the isolates from the large diphtheria outbreak in Yemen and found 13 groups, one of which (genetic cluster 217) corresponded to the predominant Yemen sublineage A.1.1, which was estimated to have originated in March 2015 (9).

**DISCUSSION**

*C. diphtheriae* was previously described as a genetically heterogeneous species (11, 18, 55). Here, we compared all available genomic sequences and defined *C. diphtheriae* sublineages, showing that they are numerous and still largely undersampled. Sublineages are separated by large genetic distances (2.4% on average), which indicates that *C. diphtheriae* is an old species that may have existed since several million years (56). The relatively rapid evolutionary rate, which was previously estimated at $\sim 1 \times 10^{-6}$ substitutions per site year$^{-1}$ within several *C. diphtheriae* sublineages (9, 16), and the frequent homologous recombination, which drives evolutionary divergence approximately five times faster than mutation (11), both contribute to the diversification of extant *C. diphtheriae* strains. This high level of genetic variation and fast...
diversification underline the need for a standard definition of \textit{C. diphtheriae} strains and sublineages based on a precise genome-based genotyping approach. Thus far, this lack has limited our understanding of the patterns of geographical spread and of the phenotypic and clinical differences among \textit{C. diphtheriae} sublineages.

Here, we developed a cgMLST genotyping approach, designed to be widely applicable across \textit{C. diphtheriae} and the closely related species, \textit{C. rouxii} and \textit{C. belfanti}. cgMLST is recognized as a highly reproducible and discriminatory approach applicable to molecular epidemiology and population biology questions (24, 57, 58). We proposed a definition of sublineages that considers the population structure, and ensured backward compatibility with MLST identifiers through their inheritance onto cgMLST-based sublineages. In the future, novel sublineage identifiers, defined systematically using the 500-cgMLST-mismatch threshold, will not be matched to MLST for practical reasons, and we propose that the cgMLST nomenclature should be prioritized over MLST. We integrated into the nomenclature, all publicly available genomes as of April 2021. The resulting genomic library is open to future community submissions and curator contributions.

Although most sublineages appear geographically restricted, sampling is currently very limited, and the current picture is strongly biased by a focus on local transmission and outbreaks. It is striking that, excluding outbreaks, no \textit{C. diphtheriae} sublineage is numerically predominant, suggesting an absence of globally successful clonal expansions. This contrasts with the situation observed in multiple bacterial pathogens where specific sublineages have come to predominate (57, 58). The population dynamics of \textit{C. diphtheriae} and its ecological drivers remains largely to be understood, but the evidence so far points to a large reservoir of diversity, strong phylogeographic structure and a lack of rapid emergence and spread of predominant sublineages. Asymptomatic carriage of \textit{C. diphtheriae} in human populations may play an important role in maintaining the large genetic pool of \textit{C. diphtheriae}.

We used a threshold of 25 allelic mismatches to define genetic clusters. This threshold captured all documented single-strain outbreak or case clusters. However, looser transmission chains of epidemiological relevance may exist, and the threshold should be considered indicative. Cryptic cluster analysis illustrates the potential value of cgMLST for uncovering previously hidden links between \textit{C. diphtheriae} infection cases. Using our definition, we noted that high diversity of genetic clusters exists within single outbreaks of classical diphtheria, such as in the former Soviet Union or in Yemen. These observations underline the diversity of transmission chains that can occur simultaneously during such events, consistent with local reservoirs of diversity and silent circulation of \textit{C. diphtheriae} until disease protection vanishes following a decrease in vaccination coverage. Strain heterogeneity during outbreaks may imply variations in toxin production or antimicrobial sensitivity, which could be highly relevant in terms of patient care and infection control priorities (9). The variation observed among subcultures of strain PW8 raises the question of a possible impact on vaccine effectiveness of the use of distinct variants of this strain in vaccine formulations (59).

The 1,305 cgMLST loci cover only approximately half of the entire genome length of a typical \textit{C. diphtheriae}. Hence, the cgMLST has an inherent limitation in capturing genetic variation. Because our main goal was classification and nomenclature, we filtered out multiple genes to prioritize genotyping robustness and call rate over strain typing resolution power. Still, our comparison of cgMLST and whole-genome SNP variation within described outbreaks shows high concordance between these methods, while illustrating that SNP calling will provide some additional resolution when needed. However, this more resolutive approach does not allow straightforward comparison of the genotypes sampled across different studies, and the two approaches should therefore be viewed as complementary.

Our phylogenetic analyses show that \textit{C. diphtheriae} genomes can acquire and lose, at small evolutionary timescales, gene regions that underlie toxigenicity and biovar. The phyletic pattern of the spuA gene presence suggests that both losses and gains may be selectively advantageous given the circumstances. Likewise, the tox gene has been acquired or lost a large number of times in the population history of \textit{C. diphtheriae}, and some of these events were traced here to a few years or decades back. Together with antimicrobial resistance and other instances of gene acquisition and loss (11, 55, 60), these
evolutionary patterns paint a picture of *C. diphtheriae* as a highly dynamic species in which sublineages may evolve at epidemiological timescales with respect to their medically important characteristics. This underlines the need to better understand the drivers of these changes and to monitor the emergence and spread of genotypes of concern, e.g., which would combine multidrug-resistant and toxigenic properties. Here, we did not look in detail at the accessory genome, and future genotype-phenotype relationship studies should be carried out to decipher the impact of *C. diphtheriae* diversity on the heterogeneity of clinical phenotypes.

Important evolutionary and epidemiological questions about *C. diphtheriae* remain unanswered, including the likelihood for tox-negative strains to turn positive during infection or short-term transmission. Asymptomatic carriage is also poorly understood, and whether *C. diphtheriae* has evolved to partially escape vaccine-induced immunity in the context of long-standing vaccination is currently unknown (59). This study provides a framework to harmonize knowledge on the infraspecific diversity of *C. diphtheriae*. The use of a unified approach to define *C. diphtheriae* strains should facilitate advances of our knowledge of the evolutionary and transmission dynamics of *C. diphtheriae* and the identification of strains with particular clinical properties or emerging behavior.

### SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

- **SUPPLEMENTAL FILE 1**, PDF file, 1.8 MB.
- **SUPPLEMENTAL FILE 2**, XLSX file, 0.1 MB.
- **SUPPLEMENTAL FILE 3**, XLSX file, 0.1 MB.
- **SUPPLEMENTAL FILE 4**, XLSX file, 0.02 MB.

### ACKNOWLEDGMENTS

We thank Melody Dazas for assistance in the early steps of the project and Annick Carmi-Leroy, Annie Landier, Nathalie Armatys, and Virginie Passet for technical assistance with the microbiological characterization and sequencing of the *C. diphtheriae* strains from the National Reference Center. We thank Vincent Enouf and the P2M core facility of Institut Pasteur for genomic sequencing. This study used computational and storage services (TARS cluster) provided by the IT Department at the Institut Pasteur, Paris, France.

To conduct the research, we used bacterial strains, which are not considered human samples. Accordingly, this research was not considered human research and is out of the scope of decree 2016-1537 (16 November 2016) implementing law 2012-300 (5 March 2012) on research involving human subjects. Therefore, no ethics approval was needed, and no informed consent was required.

All French bacteriological samples were collected in the frame of French national diphtheria surveillance and are collected, coded, shipped, managed, and analyzed according to the French National Reference Center protocols that received approval by French competent body CNIL (Commission Nationale Informatique et Libertés; approval 1474671). Other strains were obtained from culture collections.

The research conformed to the principles of the Helsinki Declaration.

We declare there are no competing interests.

This research was funded, in whole or in part, by Institut Pasteur and Santé publique France. For the purpose of open access, the authors have applied a CC-BY public copyright license to any author manuscript version arising from this submission.

S.B. conceived, designed, and coordinated the study. E.B. and S.B. coordinated the microbiological cultures of the isolates and their biochemical and molecular characterizations. J.G., M.H., A.C., and S.B. analyzed the genomic data. J.T. reviewed the clinical source data of the isolates collection. S.B. and J.G. wrote the initial version of the manuscript. All authors provided input to the manuscript and reviewed the final version.

This research was funded by institutional support from Institut Pasteur and Santé Publique France and received financial support from the French Government Investissement d’Avenir Program Laboratoire d’Excellence on Integrative Biology of Emerging Infectious Diseases (ANR-10-LABX-62-IBEID). M.H. was supported financially by a Ph.D. grant from the...
REFERENCES
1. Burkovski A. 2014. Diphtheria and its etiological agents, p 1–10. In Burkovski A (ed), Corynebacterium diphtheriae and related toxigenic species: genomics, pathogenicity, and applications. Springer, New York, NY.
2. Zasada AA, Zaleska M, Podlasiak RB, Seferynska I. 2005. The first case of septicaemia due to nontoxigenic Corynebacterium diphtheriae in Poland: case report. Ann Clin Microbiol Antimicrob 4:8. https://doi.org/10.1053/j.aml.2004.10.071-1-4-8.
3. Sharma NC, Efstratiou A, Mokrousov I, Mutreja A, Das B. 2019. Diphtheria. Nat Rev Dis Primer 5:81. https://doi.org/10.1038/s41572-019-0131-y.
4. ECDC. 2021. Diphtheria: annual epidemiological report for 2018. European Centre for Disease Prevention and Control, Solna, Sweden. https://www.ecdc.europa.eu/sites/default/files/documents/diphtheria-annual-epidemiological-report-2018.pdf.
5. Meinel DM, Kuehl R, Zbinden R, Boskova V, Garzoni C, Fadini D, Dolina M, Blümel W, Weibel T, Tschudin-Sutter S, Widmer AF, Bielicki JA, Dierig A, Heininger U, Konrad R, Berger A, Hinic V, Goldenberger D, Blach A, Stadler T, Battegay M, Sing A, Egli A. 2016. Outbreak investigation for toxigenic Corynebacterium diphtheriae wound infections in refugees from Northeast Africa and Syria in Switzerland and Germany by whole-genome sequencing. Clin Microbiol Infect 22:1003.e1–1003.e8. https://doi.org/10.1016/j.cmi.2016.08.010.
6. Scheifer C, Rolland-Debord C, Badell E, Reibel F, Aubry A, Perignon A, Patey O, Brisse S, Caumes E. 2019. Re-emergence of Corynebacterium diphtheriae. Med Mal Infect 49:463–466. https://doi.org/10.1016/j.medmal.2018.12.001.
7. Dittmann S, Wharton M, Vitek C, Ciotti M, Galazka A, Guichard S, Hardy J, Kartoglu U, Koyama S, Kreysler J, Martin B, Mercer D, Renne T, Rourke C, Steiniglas R, Strebel P, Sutter R, Trottiste M. 2000. Successful control of endemic diphtheria in the states of the former Soviet Socialist Republics: lessons learned. J Infect Dis 181(Suppl 1):S10–S22. https://doi.org/10.1086/315534.
8. Polonsky JA, Ivey M, Mazhar MKA, Rahman Z, Le Polain de Waroux O, Kato B, Jalava K, Vong S, Badjoe A, Diiz J, Finger F, Habib ZH, Cader CE, Haskew C, Kaiser L, Khan AS, Sangal L, Shirin T, Zaki QA, Salam MA, White K. 2021. Epidemiological, clinical, and public health responses to the rapid decline of diphtheria among the Rohingya population in Cox’s Bazar, Bangladesh, 2017 to 2019: a retrospective study. PLoS Med 18:e1003587. https://doi.org/10.1371/journal.pmed.1003587.
9. Badell E, Alharazi A, Criscuolo A, Almoayed KAA, Lefrancq N, Bouchez V, Guggiellini J, Hennart M, Carmi-Leroy A, Zidane N, Pascal-Perrigault M, Lebreton M, Martin V, Salje H, Toubiana J, Brisse S, Smith D, Sutcliffe IC, Titov L, Sangal V. 2017. Genomic analysis of Corynebacterium diphtheriae isolates recovered from an inner-city population demonstrates the predominance of a single molecular strain. J Clin Microbiol 55:e01581-21. https://doi.org/10.1128/JCM.01581-21.
10. Seth-Smith HM, Egli A. 2019. Whole genome sequencing for surveillance of diphtheria in low incidence settings. Front Public Health 7:235. https://doi.org/10.3389/fpubh.2019.00235.
11. Seth-Smith HM, Egli A. 2019. Whole-genome sequencing suggests transmission of Corynebacterium diphtheriae caused cutaneous diphtheria in two siblings, Germany. 2018. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull 24. https://doi.org/10.2861/j.1757-9536.2019.21800683.
12. Dangel A, Berger A, Konrad R, Bischoff H, Sing A. 2018. Geographically diverse clusters of nontoxigenic Corynebacterium diphtheriae infection, Germany, 2016–2017. Emerg Infect Dis https://doi.org/10.3201/eid2007.170206.
13. Charlton SD, Ritchie G, Lawson T, Romney MG, Lowe CF. 2019. Whole-genome sequencing of Corynebacterium diphtheriae isolates recovered from an inner-city population demonstrates the predominance of a single molecular strain. J Clin Microbiol 58:e001651-19. https://doi.org/10.1128/JCM.01651-19.
14. Pivott D, Fanto A, Badell-Ocando E, Benouachouk M, Astruc K, et al. 2019. Carriage of a single strain of nontoxigenic Corynebacterium diphtheriae biovar Belfanti (Corynebacterium belfanti) in four patients with cystic fibrosis. J Clin Microbiol 58:e00049-20. https://doi.org/10.1128/JCM.00049-20.
15. Xiaoli L, Benoilel J, Peng Y, Ankele J, Cassiday PK, Kay M, McKerimn S, Duchin JS, Kawakami V, Lindquist S, Acosta AM, DeBolt C, Tondella M, Weigand MR. 2020. Genomic epidemiology of nontoxigenic Corynebacterium diphtheriae from King County, Washington State, USA between July 2018 and May 2019. Microb Genom 6. https://doi.org/10.1099/mgen.0.000467.
16. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial nucleotide sequence analysis. Nat Rev Microbiol 11:728–736. https://doi.org/10.1038/nrmicro3093.
17. Du Plessis M, Wolter N, Allam M, de Gouveia L, Moosa F, Nithoe G, Blumberg L, Cohen C, Smith M, Mutuwedzi P, Thomas J, Horne V, Moodley P, Archy M, Mahabeer Y, Mahomed S, Kuhn W, Miliana K, McCarthy K, von Gottberg A. 2017. Molecular characterization of Corynebacterium diphtheriae outbreak isolates, South Africa, March-June 2015. Emerg Infect Dis 23:1308–1315. https://doi.org/10.3201/eid2308.160239.
18. Schaeffer J, Huhulescu S, Stoeger A, Allerberger F, Ruppitsch W. 2020. Assessing the genetic diversity of Austrian Corynebacterium diphtheriae clinical isolates, 2011–2019. J Clin Microbiol 58:e02529-20. https://doi.org/10.1128/JCM.02529-20.
19. Hoefer A, Pampaka D, Herrera-León S, Peišov S, Varona S, et al. 2020. Molecular and epidemiological characterization of toxigenic and nontoxigenic C. diphtheriae, C. belfanti, and C. ulcerans isolates identified in Spain from 2014 to 2019. J Clin Microbiol 58:e02410-20. https://doi.org/10.1128/JCM.02410-20.
20. Will RC, Ramamurthy T, Sharma NC, Veeraraghavan B, Sangal L, Haldar P, Pragasam AK, Vasudevan K, Kumar D, Das B, Heinz E, Melnikov B, Baker S, Sangal V, Dougan G, Mutreja A. 2021. Spatiotemporal persistence of multiple, diverse clades and toxins of Corynebacterium diphtheriae. Nat Commun 12:1500. https://doi.org/10.1038/s41467-021-21870-5.
21. Criscuolo A, Brisse S. 2013. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. Genomics 102:500–506. https://doi.org/10.1016/j.ygeno.2013.07.011.
22. Liu Y, Schröder J, Schmidt B, 2013. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics 29:308–313. https://doi.org/10.1093/bioinformatics/btt690.
23. Maqoq T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads. Bioinformatics 27:2597–2598. https://doi.org/10.1093/bioinformatics/btr413.
