scientific reports

Check for updates

OPEN

Association of pathogenic determinants of *Fusobacterium necrophorum* with bacteremia, and Lemierre's syndrome

Alessia Carrara[1], Claire Bertelli[1], Céline Gardiol[2], Bastian Marquis[1], Diego O. Andrey[3], Jacques Schrenzel[3], Trestan Pillonel[1] & Gilbert Greub[1,2]✉

Fusobacterium necrophorum **is a Gram-negative anaerobic bacterium responsible for localized infections of the oropharynx that can evolve into bacteremia and/or septic thrombophlebitis of the jugular vein or peritonsillar vein, called Lemierre's syndrome. To identify microbial genetic determinants associated with the severity of this life-threatening disease, 70** ***F. necrophorum*** **strains were collected and grouped into two categories according to the clinical presentation: (i) localized infection, (ii) bacteremia with/without Lemierre's syndrome. Comparative genomic analyses revealed two clades with distinct genetic content, one clade being significantly enriched with isolates from subjects with bacteremia. To identify genetic determinants contributing to** ***F. necrophorum*** **pathogenicity, genomic islands and virulence factor orthogroups (OVFs) were predicted. The presence/absence profiles of OVFs did not group isolates according to their clinical category, but rather according to their phylogeny. However, a variant of** ***lktA*****, a key virulence factor, with a frameshift deletion that results in two open reading frames, was associated with bacteremia. Moreover, a genome-wide association study identified three orthogroups associated with bacteremic strains: (i) cas8a1, (ii) a sodium/solute symporter, and (iii) a POP1 domain-containing protein. Further studies must be performed to assess the functional impact of** ***lktA*** **mutation and of these orthogroups on the physiopathological mechanisms of** ***F. necrophorum*** **infections.**

Keywords Lemierre's syndrome, Anaerobes, Thrombophlebitis, Sepsis, Bacteremia, Tonsillitis, Fusobacterium

Abbreviations

OVF	Orthogroups of virulence factors
LS	Bacteremia with Lemierre's syndrome
BAC	Bacteremia without Lemierre's syndrome
LI	Localized infection
PBC	Positive blood culture
PSTS	Persistent sore throat syndrome
ANI	Average nucleotide identity
SNP	Single-nucleotide polymorphism
CDS	Coding Sequence
ORF	Open reading frameF
HGT	Horizontal gene transfer
GWAs	Genome-wide association study
GI	Genomic island

Fusobacterium necrophorum is a Gram-negative, non-motile, non-spore-forming, and obligate anaerobe bacterium. Among the two subspecies of *F. necrophorum* identified, subsp. *necrophorum* is uniquely associated with

[1]Institute of Microbiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland. [2]Service of Infectious Diseases, Department of Medicine, Lausanne University Hospital, Lausanne, Switzerland. [3]Service of Bacteriology and Infectious Diseases, Department of Medicine, University Hospital of Geneva, Geneva, Switzerland. ✉email: gilbert.greub@chuv.ch

animal infections, whereas subsp. *funduliforme* is responsible for both animal and human infections[1,2]. *F. necrophorum* is one of the main players in oropharyngeal infections, retropharyngeal abscesses, tonsillitis, persistent sore throat syndrome (PSTS), and post-surgery mediastinitis. However, it can also cause otogenic, odontogenic, and deep neck space infections[3]. More severe clinical presentations may occur when the bacteria reach the jugular or peritonsillar vein. Indeed, the most common complication of *F. necrophorum* infection, called Lemierre's syndrome, is characterized by a thrombophlebitis of the internal jugular vein and/or of the peritonsillar vein that leads to secondary metastatic pyogenic infections and bacteremia[4,5]. Occasionally, bacteremia is also observed without clinical progression toward Lemierre's syndrome, i.e. in absence of thrombophlebitis.

Although *F. necrophorum* is considered as a commensal of the gastrointestinal and genitourinary tract of animals such as cattle and humans, its classification as a commensal of the human upper respiratory tract in the oropharyngeal cavity is still debated. Indeed, its presence was detected in two studies with a prevalence of 0 and 21% among 100 and 92 healthy control throat swabs, respectively[6,7]. This open question highlights our limited understanding of the pathogenesis of *F. necrophorum* infection in humans. The current hypothes is suggests that the pathogen could be acquired through eating or drinking contaminated food and might cause infection in individuals with mucosal lesions and/or poor oral hygiene[8]. In favorable conditions, the bacterium would hence colonize the oral cavity as well as the nasopharynx before spreading to the adjacent tissues and to the vein through mechanisms that remain poorly understood.

The factors facilitating the invasion of *F. necrophorum* in submucosal tissues are (i) the coinfection with Epstein-Barr virus, with group A β-hemolytic streptococcus, and with other aerobic or anaerobic bacteria, (ii) the presence of surrounding antibiotic-resistant bacteria able to provide a protective advantage to *F. necrophorum*, commonly susceptible to antibiotics, and (iii) host characteristics[9]. Indeed, Lemierre's syndrome is predominant in young adults (14–25 years old) with a similar ratio of males and females[10,11]. One case per million is reported per year worldwide[12]. However, its incidence has been increasing in the last few years, potentially because of a reduction of tonsillectomies and fewer antibiotics prescriptions, but also thanks to the improvement in laboratory techniques to isolate and identify anaerobic microorganisms[13].

Several studies have been conducted to characterize *F. necrophorum* subsp. *necrophorum* infections in animals and to detect important virulence factors given their impact on the cattle industry, raising major economic concerns[14]. Adhesins, endotoxins, hemolysins, leukotoxins, and hemagglutinins have been identified to mediate *F. necrophorum* pathogenesis. However, despite the ability of the leukotoxin to trigger necrosis at high concentrations and its association with the more virulent subsp. *necrophorum*[15], the mode of action of this toxin remains unknown. Additionally, the FadA adhesin, which mediates the attachment to the host cells, has been characterized in the subsp. *necrophorum* suggesting this bacterium to be an active invader of epithelial cells[16]. Conversely, the lack of the gene has been reported in subsp. *funduliforme*.

Given the many open questions on *F. necrophorum* pathogenesis in humans and its main drivers, we initiated a multicentric retrospective study to assess whether specific genetic determinants, in particular virulence factors, are associated with specific clinical presentations in humans. Furthermore, this study provides a precious database of *F. necrophorum* genomes of clinical relevance.

Methods and materials

Strain collection and metadata

This retrospective multicentric study collected 26 F. necrophorum strains isolated from blood culture and stored routinely since 1999 and from other sample types since 2011–2012 onwards at the University Hospital of Lausanne (CHUV), and at the University Hospital of Geneva (HUG). . Uniform metadata regarding clinical information, patient demographics and comorbidities, laboratory and microbiological analyses, blood culture results, and infection localization were acquired retrospectively. The study exclusively incorporated patients with confirmed diagnoses, excluding isolates from individuals with asymptomatic infections. Additionally, to increase the sample size, forty-four genomes and the related metadata containing the patient clinical presentation were retrieved from the NCBI database (BioProject ids: PRJNA354964, PRJNA315619, PRJNA824050) and the corresponding literature (see Supplementary File 1, Table S2). All isolates were grouped into three categories according to the clinical presentation: localized infection (LI), bacteremia without Lemierre's syndrome (BAC), and bacteremia with Lemierre's syndrome (LS). All samples from Lausanne and Geneva included in this study are described in Supplementary File 2.

Localized infection included cases uniquely characterized by oropharyngeal and/or otological infections. Bacteremia without Lemierre's syndrome included cases where the infection reached the bloodstream as documented through a positive blood culture or causing sepsis and/or metastatic infections; with metastatic infection, we refer to a secondary and distal infection compared to the initial site. Lemierre's syndrome category included cases with a documented diagnosis of Lemierre's syndrome, characterized by a primary bacterial oropharyngeal infection, followed by thrombophlebitis of the internal jugular vein and/or peritonsillar vein that ultimately leads to secondary metastatic pyogenic infections and bacteremia[4,5].

For many analyses, BAC and LS categories were considered together in the positive blood culture (PBC) group.

All methods were carried out in accordance with relevant guidelines and regulations related to good laboratory practices and biosafety rules; the strains were isolated from accredited laboratories. All human clinical data were collected in accordance with the ethical standards set by the regional, national Swiss research committees, as well as international rules. The research has been carried out solely on bacterial strains and clinical data, without any direct further analysis of patient samples or question to the patients. All experimental protocols were accepted by the president of the Canton ethical committee (CER-VD) on 16/02/2012".

DNA extraction, and whole genome sequencing

The bacterial DNA was extracted from clonal cultures using the Wizard SV Genomic DNA Purification System (Promega, Madison, USA). The library was prepared with Nextera XT DNA Sample Preparation kit and further sequenced by 150 bp paired-end reads on a MiSeq (Illumina, San Diego, CA). In addition, sample GEN_281 was also sequenced with Pacific BioSciences (PacBio) technology to increase the assembly quality.

Genome assembly and annotation

Read quality was assessed using FastQC version 0.11.08 (Andrews S. (2010), http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Trimmomatic version 0.32[17] was used to filter low-quality reads and reads shorter than 150 bp. Reads were assembled with SPAdes genome assembler version 3.15.4[18] using k-mer sizes ranging from 43 to 127 bases. Quast version 3.1[19] was used to select the best assembly based on the lowest number of contigs and largest N50, the shortest contig length that needs to be included to cover 50% of the genome. Contigs shorter than 1000 bp and low k-mer coverage contigs (<2x) were discarded from the assemblies. The remaining contigs were reordered based on the RefSeq reference genome of strain F1260 using Mauve Contig Mover[20]. Genes were annotated using Bakta version 1.4.0[21]. Pyani version 0.2.11[22] was used to compute the average nucleotide identity (ANI) and ensure that all genomes were members of *F. necrophorum* subsp. *funduliforme*. A long and short-read hybrid assembly was performed for sample GEN_281 using Unicycler version 0.4.9[23] with default settings.

Comparative genomics database and virulence factor identification

The comparative genomic tool zDB version 1.0.5[24] was used to build a database containing orthology and phylogenetic inferences, as well as functional annotations retrieved against the COG database[25], and the Pfam database[26], for the 70 *F. necrophorum* genomes. FastTree version 2.1.11[27] was used to build a phylogenetic tree based on concatenated protein sequence alignments of 1461 single-copy core orthologs identified by OrthoFinder version 2.5.2[28]. iTOL v6.8.2[29] was used to visualize and annotate the phylogeny. The pam function from the cluster package version 2.1.2[30] was used to identify clusters of genomes based on orthogroup presence/absence. The optimum number of clusters was determined using the maximal Calinski-Harabasz (CH) Index computed using index.G1 function from the clusterSim package version 0.51-3[31].

Five virulence factor databases (Victors[32], VFDB[33], PHI-base[34], SWISS-PROT[35] and PATRIC[36]) and an ad-hoc *F. necrophorum*-database were used to identify virulence factors genes with BLASTP[37]. The ad-hoc database containing 108 unique gene sequences was generated searching in UniProt for proteins annotated with the word "*Fusobacterium necrophorum*" in association with one of the following words describing virulence factors reported in literature: *virulence, ecotin, LPS, FadA, neutrophil-activating protein A, lktA, lktB,* and *lktC* (see Supplementary File 1, Table S3). BLASTP hits with more than 40% of identity, as well as ≥60% of query and subject coverage were considered candidate virulence factors. All members of the orthogroups encoding for a VF were further used in subsequent VF analyses and visualization. Multiple sequence alignment of *lktA* gene was performed using Mafft version 7.475[38]. The phylogenetic tree reconstruction was obtained with FastTree version 2.1.10[27], and its visualization was performed with phylo.io[39]. BLASTN and TBLASTN[37] were used to evaluate the sequence conservation of the genes of *lkt* operon and the promoter. beta.phylo.io[39] was used to compare the topology of the dendrogram constructed using Euclidean distances computed on OVF presence/absence and the phylogenetic tree built from the concatenated protein sequence alignments of single-copy core orthologs.

Genomic islands and genome recombination

IslandCompare[40] and PHASTEST (annotation mode: deep)[41] were used to identify genomic islands and prophage regions, respectively. The output of the two tools was merged and clusters of genomic islands were computed applying MCL algorithm[42] on their mash distances (Sketch individual sequences = True, k-mer size= 16, inflation= 5)[43]. ChromoPlot[44] was used to visualize the genomic island clusters along the genomes. The recombination events were assessed using Gubbins version 2.4.1[45]. The BWA-MEM, integrated into Snippy version 4.6.0 (available at https://github.com/tseemann/snippy), was used to compute the core genome alignment of all genomes against the reference F1291. This alignment was then provided as input to Gubbins.

Genome-wide association study (GWAS)

A genome-wide association study was performed to identify any genomic signature associated with the clinical category using the linear mixed models of Pyseer version 1.3.10[46] (*–similarity* phylogenetic distances). The features evaluated were (i) the presence/absence of orthogroups, (ii) genomic island clusters, (iii) the k-mer composition computed using fms-lite (available online at: https://github.com/nvalimak/fsm-lite), (iv) and the SNPs identified with Snippy version 4.6.0 (snippy-core) (available online at https://github.com/tseemann/snippy)) against genome F1291 as reference. Synonymous SNPs and SNPs located outside coding sequences, as predicted by SnpEff[47], were discarded from the analysis and filtered out with SnpSift[48]. In Pyseer, all features present in less than 10% or present in all genomes were discarded. All evaluations were corrected for multiple testing with Bonferroni correction (alpha-error<0.05). To further identify whether the presence/absence of orthogroups, genomic island clusters, or SNPs was relevant to distinguish genomes associated with the clinical categories, Boruta function from the R package Boruta version 8.0.0[49] was used.

Statistical analyses

The statistical tests were performed using the R package stats (version 4.3.1)[50] and R package Barnard (version 1.8)[51]. Continuous variables collected in the clinical metadata were described as median with interquartile range (IQR) and compared across groups using a Wilcoxon-test, while for categorical binary variables, the Chi-squared

test followed by Yates continuity correction was computed. Wilcoxon-test was applied to test the number of virulence factors detected per clinical category. Instead, Bernard's test was computed to compare the ratio of BAC+LS/LI genomes between the two clades, and the ratio of BAC+LS/LI genomes between the ones carrying the *lktA* type 1b variant and those without. The Chi-squared test followed by Yates continuity correction was used to evaluate the significance of the relevant genomic features identified by the Boruta algorithm[49].

Results

Cohort description

The study cohort consisted of 70 *F. necrophorum* strains, including 44 publicly available genomes and 26 newly sequenced strains isolated from patients admitted to two university hospital centers in Switzerland (Table 1). To classify the various clinical presentations associated with *F. necrophorum* infections, physicians retrospectively grouped the patients into three clinical categories based on the collected clinical metadata. Eleven cases were classified as "bacteremia" due to the presence of positive blood culture and/or a diagnosis of Lemierre's syndrome (PBC group). Among these, 3 patients were diagnosed with Lemierre's syndrome (LS category), whereas the other 8 did not exhibit disease progression into thrombophlebitis or metastatic thrombotic infections (BAC category). The remaining 15 cases were categorized as localized infections only (LI category).

The patients, of whom 14 were males and 12 were females, were on average 28 years old (sd. 14.26). Bacteremic patients diagnosed with Lemierre's syndrome had a significantly lower age compared to others. Almost half (n=12) of the patients presented polymicrobial infections, including *F. necrophorum*. Among these, 9 belonged to the LI category. No patient showed hematological malignancy, oncological diseases, or alcohol use disorders, but one patient was immunosuppressed due to corticosteroid treatment.

The inclusion of publicly available genomes increased the number of strains classified in the LS category from 3 to 11, and those assigned to the LI category from 15 to 51. Ultimately, the cohort consisted of 11, 8, and 51 strains in the LS, BAC, and LI categories, respectively.

Two main clades share distinct orthogroups

The genomes, with a size ranging from 1.98 to 2.45 Mbp, exhibited a GC content of 35% and a coding density of 92%. A total of 3228 orthogroups were identified in the dataset of 70 *F. necrophorum* strains. Among those, 1461 orthogroups (45%) composed a large and stable core genome, while the accessory orthogroups were most uniquely present in a single or a few genomes, as can be seen by the distribution of orthogroup conservation in red (Fig. 1a). Furthermore, the number of orthogroups rising with the increasing number of genomes is representative of an open pangenome, suggesting that *F. necrophorum* subsp. *funduliforme* is genetically diverse and affected by a significant number of gene acquisitions and gene losses.

The Pam-clustering based on orthogroup presence-absence identified two main genomic clusters (Fig. 1b; Supplementary File 1, Figure S1). While 143 orthologues were uniquely present in cluster 1 which contained 17 genomes and shared 2597 orthogroups, 406 orthogroups were uniquely identified in cluster, 1, which was made of 53 members and 2765 shared orthogroups. The classification of the genomes in these clusters mostly reflected their position along the phylogenetic tree built on the concatenated alignments of single-copy core orthologs (Supplementary File 1, Figure S2). However, genomes BRON_6, BRON_13, and GEN_298 displayed deeper branching in the phylogenetic tree suggesting that further clades will likely be distinguished with increased sample size and diversity. Indeed, while the average nucleotide identity between the cluster 1 and 2, was 0.996 and 0.995, respectively, and the average nucleotide identity between the clusters was 0.991, the average nucleotide

	PBC			p-value	
	LI	BAC	LS	LI vs. PBC	BAC vs. LS
Demographics					
Number of patients	15	8	3		
Sex (M/F)*	11/4	2/6	1/2	0.053	1
Age (yr)†	29 [0–67.3]	28 [10–45.1]	19 [**15.3**–21.3]	0.435	**0.018**
Clinical information					
Sepsis (yes)*	1	3	3	**0.023**	0.24
Metastatic infection (yes)*	1	2	3	0.065	0.122
ICU stay (days)†	0 [0–3.1]	0 [0–10]	5 [0–10]	0.566	0.7
Hospital stay (days)†	4.5 [0–32.4]	5.5 [0–23.3]	32 [0–71.8]	0.361	0.183
Microbiological analyses					
Positive blood cultures (yes)*	0	8	3	<**0.001**	–
Co-pathogen (yes)*	9	2	1	0.151	1

Table 1. Patient demographics according to the clinical presentation. Demographic data are presented as numbers or median with IQR in squared brackets according to the associated clinical category (LI: localized infection, BAC: bacteremia without Lemierre's syndrome, LS: bacteremia with Lemierre's syndrome, PBC: positive blood culture group, which includes both BAC and LS genomes). The statistical tests are performed using the *Chi-squared test followed by Yates continuity correction for categorical variables, and the † Wilcoxon test for continuous variables. Significant values are in bold.

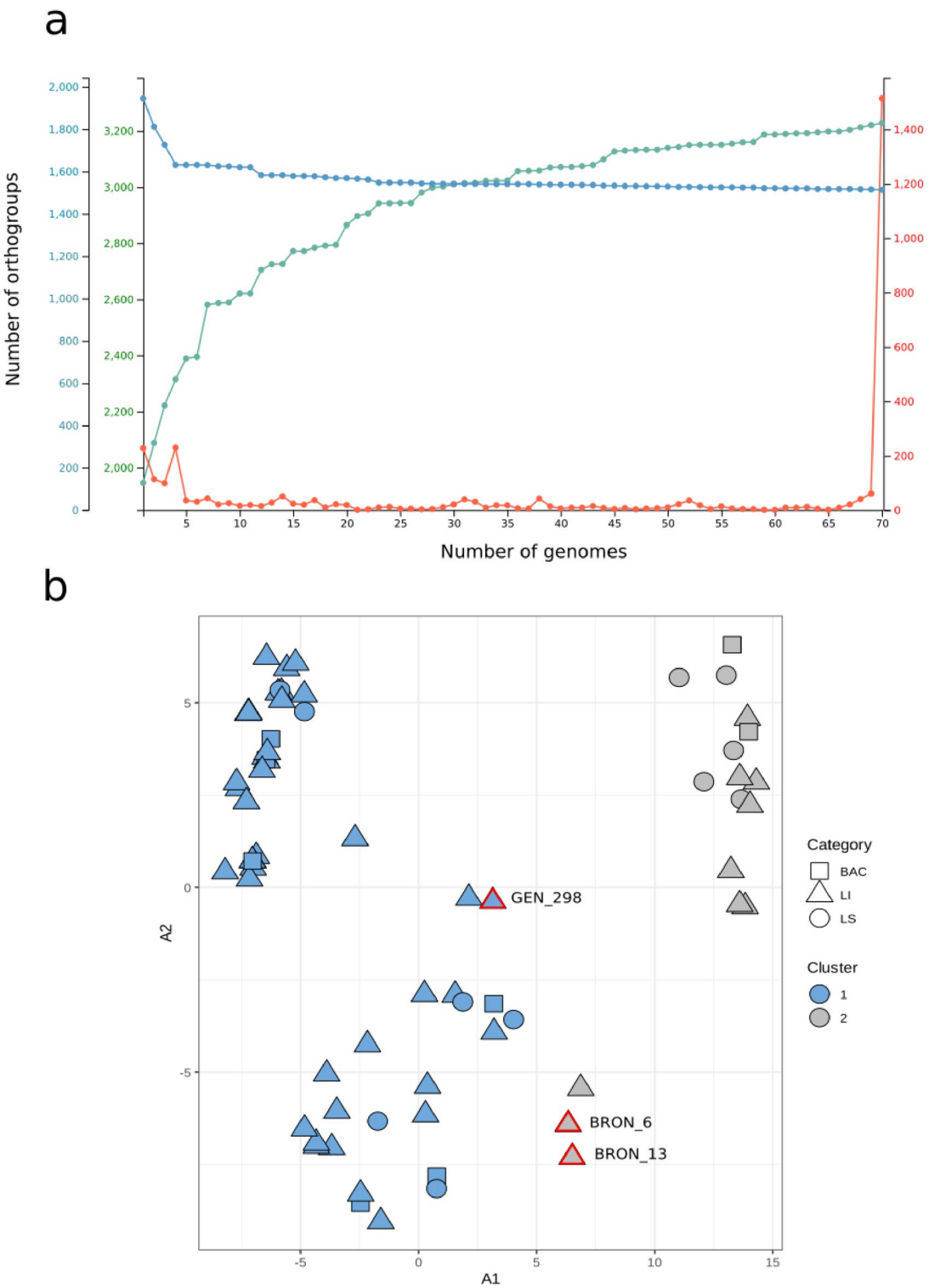

Fig. 1. Overview of the genetic diversity. (**a**) Accumulation/rarefaction plots of the orthogroups retrieved in the complete dataset composed of 70 *F. necrophorum* genomes. The red curve represents the number of orthogroups only present in a subset of n genomes, from 1 to 70. Around 50% of the orthogroups are conserved in all genomes. The blue curve indicates the number of shared orthologs when comparing 2 to 70 genomes, hence representing the core genome. The green curve indicates the total number of orthogroups with the increasing number of genomes. (**b**) Pam clustering built based on Euclidean distances out of orthogroups presence/absence. Cluster 1 contains genomes of Clade B, and GEN_298; while the member of Cluster 2 are the genomes of Clade A, BRON_6, and BRON_13.

identity between GEN_298 and cluster 1, and between BRON_6-BRON_13 and cluster 2 decreased to 0.982 and 0.986, respectively (Supplementary File 1, Figure S3). Thus, we decided to set aside BRON_6, BRON_13, and GEN_298 and considered two refined clades presenting unique genetic content and significant evolutionary divergence (Supplementary File 1, Figure S2 and Figure S4); Clade A, composed of 15 members and characterized by a unique set of 102 orthogroup, and Clade B, composed of 52 members and distinguished by 398 unique orthogroups (Fig. 3a). Although, each clade included genomes classified in all three clinical categories, Clade A comprised 50% of bacteremic strains, while the larger Clade B comprised 75% of isolates from localized infections. The ratio of BAC and LS over LI cases resulted significantly higher in Clade A as compared to Clade B (Bernard's test, p-value=0.04) (Supplementary File 1, Table S1).

Most virulence factors are shared by all genomes

To characterize the virulence profile of *F. necrophorum* genomes, orthogroups of virulence factors (OVFs) were identified and tested for association with the clinical presentations. On average 288 OVFs were retrieved per genome (sd 2.36). Of the identified OVFs, 87% were shared by the entire collection of genomes, and only 18 were present in less than half of them (Fig. 3b). OVFs encoding for ecotin, hemolysin D, the toxin component of the toxin-antitoxin system of the Fic family, cold shock protein CspB and CspC, lipopolysaccharide biosynthesis protein, OmpH, and cell envelope integrity protein CreD were ubiquitously present and appeared to be highly conserved (>98% protein sequence identity). A total of 42 OVFs were identified to be present in at least one but not all genomes. However, the dendrogram built on the Euclidean distances computed on the OVFs presence/absence mostly reflected the core-genome phylogeny, rather than the clustering according to the clinical categories (Supplementary File 1, Figure S5 and Figure S6). This suggested that VF gain/loss occurred in the common ancestor of each clade. Likewise, although OVFs generally showed high amino-acid sequence conservation across all genomes, in some cases they diverged according to the clades (Supplementary File 1, Figure S7).

No difference in the number of single-copy OVFs and multiple-copy OVFs was identified between the genomes of Clade A and Clade B, or between the genomes of the three clinical categories (Supplementary File 1, Figure S8). No OVF characterized the positive blood culture (PBC) as opposed to the LI strains. However, 6 OVFs were found to be associated with a subset of 9 to 1 strain in the localized-infection category. They encoded for type II toxin-antitoxin system RelE/ParE family toxin, adhesin, iron chelate uptake ABC transporter, cas2, and 4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase (Fig. 3a).

A frameshift-causing deletion of lktA gene is associated with positive blood culture

The *lkt* operon is a primary virulence factor of *F. necrophorum* infections composed of *lktB*, *lktA*, and *lktC* genes[52]. Recognizing the significance of the lkt virulence factor in *F. necrophorum* pathogenicity, we assessed the sequence conservation of *lkt* operon and its promoter. Divergence in nucleotide and amino acid sequences of these genes among the genomes was observed, although a high similarity was maintained within each phylogenetic clade. Indeed, variations in the amino acid sequence of 1%, 30%, and 5% were observed for *lktB*, *lktA*, and *lktC*, respectively, between the two different clades. Additionally, slight sequence variations between clades were identified in the promoter region of the *lktABC* operon (Supplementary File 1, Figure S9).

The multiple sequence alignment of both nucleotide and amino acid sequence of *lktA* reflected the phylogenetic tree obtained from the concatenated alignments of single-copy core orthologs (Fig. 4c). However, all genomes of Clade B presented an open reading frame (ORF) annotated as *lktA* with a length ranging from 8781 to 9690 bp. A deletion of 890 nt at position 5581 was identified as the major genomic difference between the shortest and the longest gene variant. Instead, two ORFs were detected in the region of *lktA* gene in 14 out of 15 genomes of Clade A. The first and the second ORF shared 72.6% and 76.8% of amino acid sequence similarity with the LktA sequence of F1260, a representative member of Clade B. This truncated variant carried a deletion of two nucleotides in positions 3918-3919 of *lktA* of F1260 that caused a shift in the translational reading frame, resulting in a premature stop codon (Fig. 4a, b). This variant, named *lktA* type 1b[53], showed a significant association with bacteremic strains (Bernard's test, p-value= 0.035).

Genomic island distribution and genome recombination

Since genomic islands (GI), which are regions acquired through horizontal gene transfer (HGT), often confer an adaptative advantage to the recipient strain[54], we tested whether any GI was associated with the pathogenicity of *F. necrophorum*. Therefore, we grouped clusters of genomic islands by sequence similarity and evaluated their association with different clinical presentations. Among the 107 genomic island clusters, 58% were unique to a single genome in the dataset. Clade A and Clade B each harbored 26 and 86 genomic island clusters, with 12 and 72 clusters, respectively, unique to each clade. The distribution of genomic island clusters reflected the phylogeny rather than the clustering according to the clinical presentations, as observed for orthogroups of virulence factors (Fig. 2b). No genomic island clusters were ubiquitously present across the members of any clinical category. A single cluster was uniquely detected in the LS category, but only in 2 genomes, while 14 clusters were uniquely present in LI categories occurring in either 2, 3, or 7 genomes out of 51.

Recombination is a fundamental mechanism underlying HGT allowing the generation of genetic diversity. Furthermore, recombination events may impact the reconstruction of bacterial evolution hiding the signal of putative point mutations outside the recombinant regions. To assess whether the phylogenetic distance between genomes associated with similar clinical presentation was impacted by recombination, we predicted recombination events and built a phylogenetic tree correcting for those. Gubbins[45] predicted multiple recombination events occurring on the internal branches of the two main clades, demonstrating that the two clades diverged enough to have their own unique sets of recombinant regions shared by multiple strains with common descent (Supplementary File 1, Figure S10). While, BRON_6, GEN_298, and BRON_13 shared recombinant regions

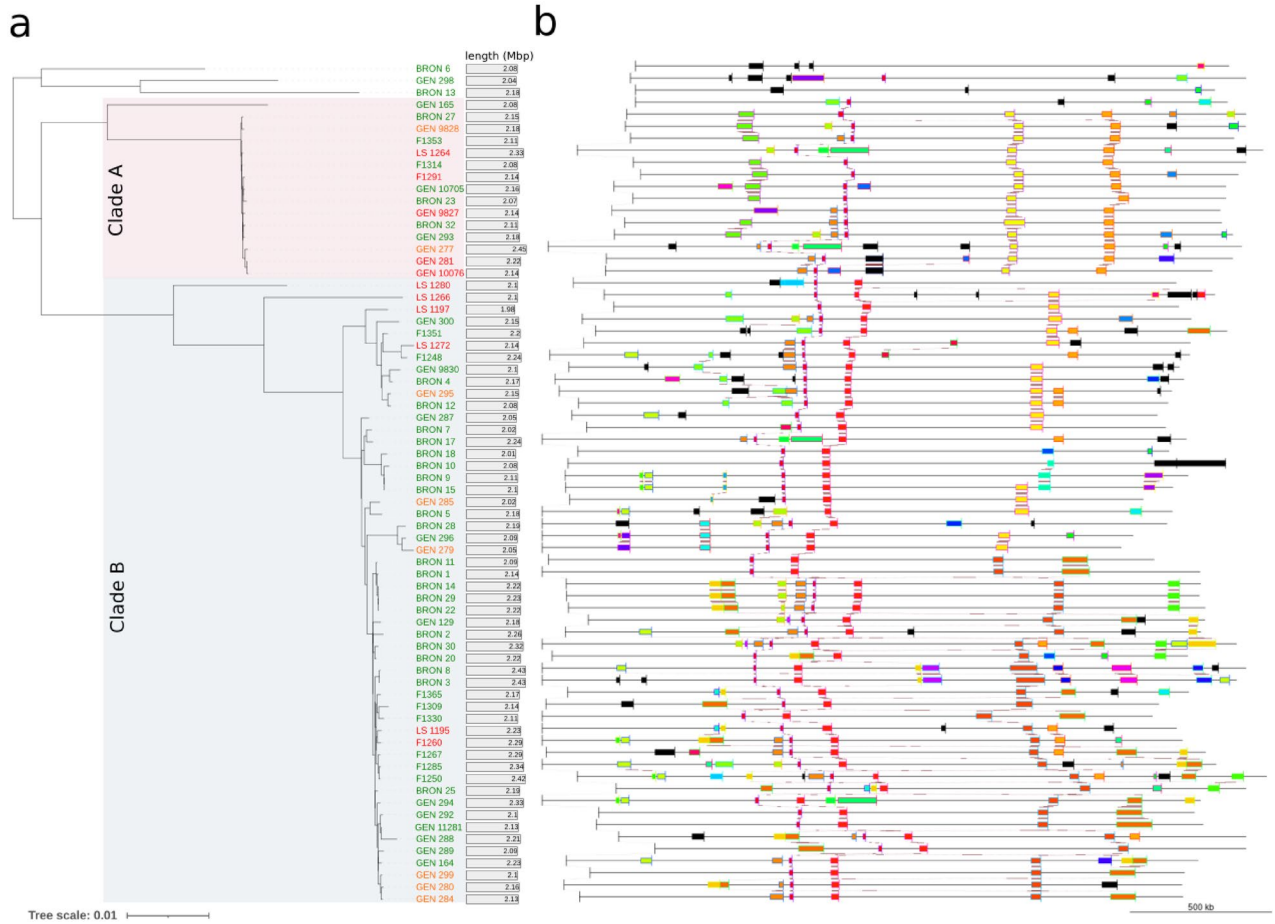

Fig. 2. Phylogenetic reconstruction and mobile elements. (**a**) Midpoint rooted phylogenetic tree based on the concatenated alignments of single-copy core orthologs. The genomes are colored according to the assigned clinical category (red: LS, orange: BAC, green: LI), and the two main clades are highlighted in light blue and pink for Clade A and Clade B, respectively. (**b**) Distribution of clusters of similar genomic islands along the genomes. Clusters with low or no similarity to any of the identified horizontally transferred regions are shown in black.

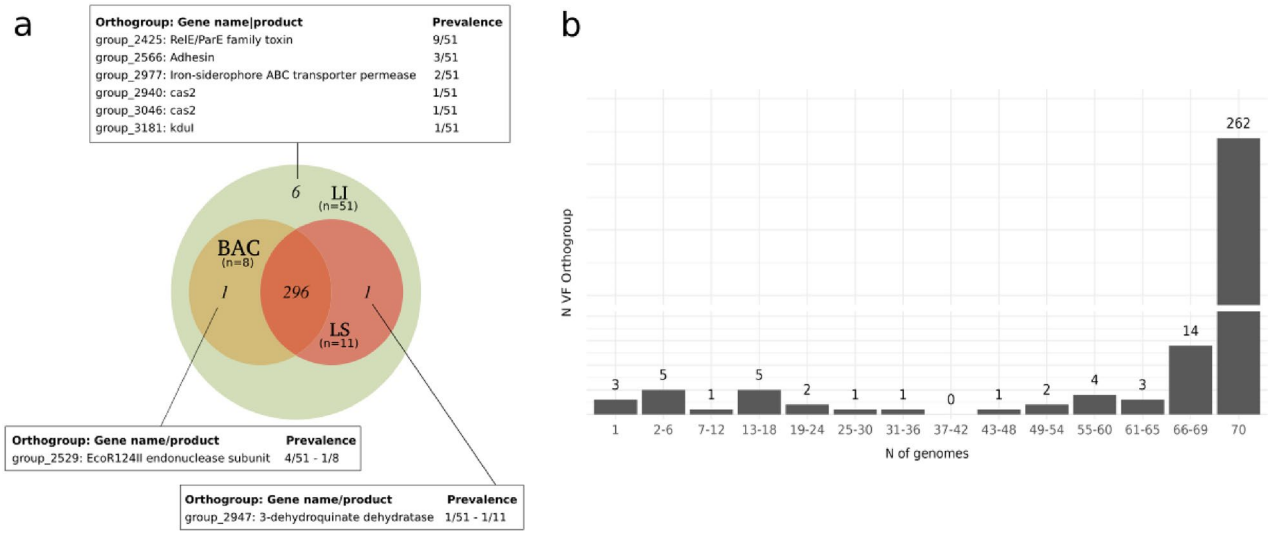

Fig. 3. Distribution of orthogroups of virulence factors (OVFs). (**a**) Venn Diagram representing the number of virulence factor orthogroups shared between the clinical categories. LS, BAC, and LI categories shared 296 OVFs, 6 OVFs were uniquely present in LI. (**b**) Distribution of OVFs present in a subset of n genomes, from 1 to 70. Among 304 OVFs, 262 were ubiquitously present in the entire *F. necrophorum* dataset.

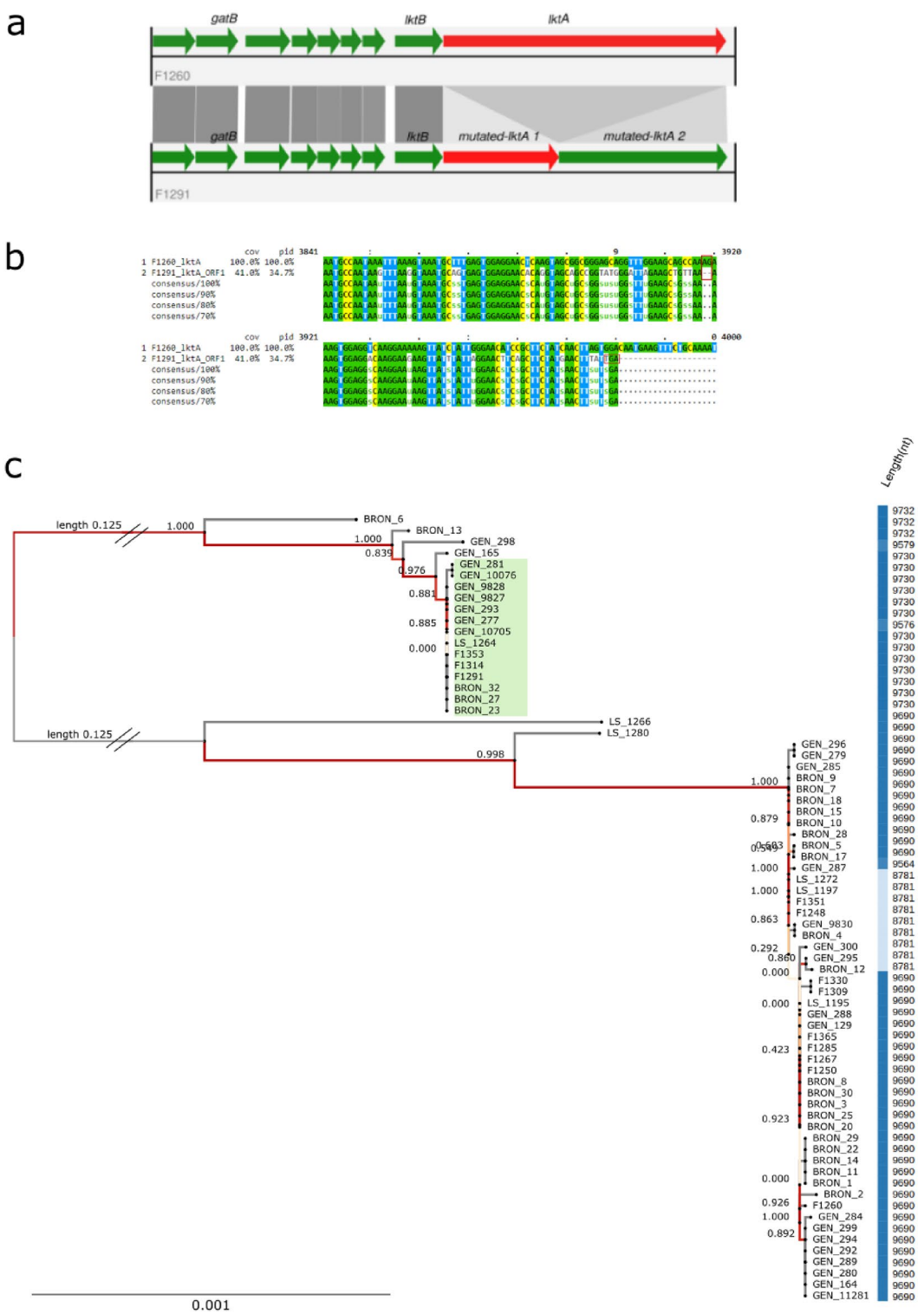

Fig. 4. Evaluation of *lktA* gene variants and *lktA* phylogeny. (**a**) Alignment of the predicted CDSs of the *lkt* region of F1260 of Clade B and the *lktA* region of F1291 of Clade A. Along F1291 two loci were identified as *lktA*-like genes (Mutated *lktA* 1 and Mutated *lktA* 2). (**b**) Pairwise nucleotide sequence alignment of *lktA* F1260 and ORF1 of *lktA* F1291. The deletion of two nucleotides in position 3918 and 3919, and the downstream TGA stop codon are displayed in the second line of the alignment corresponding to ORF1 of *lktA* F1291. (**c**) Phylogenetic tree obtained from the multiple sequence alignment of *lktA* nucleotide sequences. The length of *lktA* gene, or the full region where lktA ORF1 and ORF2 are predicted, were considered. Highlighted in green those genomes presenting the deletion in the *lktA* region that results in two predicted ORFs. The bootstrap values are reported along the phylogeny and the branches are colored according to them.

within themselves and with the members of Clade A, GEN_165 shared a few ones with Clade B. Recombinations occurring on terminal branches were mainly identified in BRON_6, LS1280, and LS1266. However, the phylogenetic tree constructed on the putative point mutations outside of the recombined regions showed high similarity with the phylogenetic tree based on the concatenated alignments of single-copy core orthologs. Therefore, the reconstruction of the species tree based on concatenated single-copy orthologs was not significantly impacted by conflicting phylogenetic signals resulting from recombination events. Indeed, closely related genomes within clade A and clade B maintained their clonal identity. Instead, the evolutionary relationship of genomes located on deeper branches, as GEN_165, BRON_6, GEN_298, and BRON_13, requires more similar strains to be solved since it might be negatively impacted by the potential long-branch attraction occurring among them[55] and the limited resolution in the recombination especially along GEN_165.

Genome-wide association with clinical categories

Genome-wide association study (GWAS) is a powerful methodology to identify features statistically associated with a phenotype of interest among a large dataset. Here, we aimed to identify potential features associated or not with Lemierre's syndrome among all bacteremic strains (LS vs. BAC category), and with the dissemination of the infection in the bloodstream in comparison to localized infections (PBC vs. LI category). The features tested were: (i) the presence/absence of orthogroups that may provide a specific function, (ii) the presence/absence of clusters of similar genomic islands, whose acquisition may provide ecological advantages, as described above, (iii) the k-mers which highlight both short sequence variations and gene presence/absence, and (iv) the presence/absence of SNPs that reflect single point mutations.

In both evaluations, Pyseer[46] did not identify any feature as significantly associated with the clinical presentations (alpha-error<0.05). However, Boruta[49], the feature selection algorithm that works as a wrapper around Random Forest, identified 7 orthogroups and 3 clusters of genomic islands as relevant features for the given predictions. Indeed, 6 orthogroups, not classified as OVFs, were evaluated as important in distinguishing PBC-associated from LI-associated genomes (Table 2). Five of them were additionally supported by a significant Chi-squared test. Group_13, which encodes for DUF1353 domain-containing protein, and group_1955, which encodes for signal recognition particle-docking protein FtsY were more prevalent in the localized infection category. The homologs of these orthogroups were located on multiple and different clusters of genomic islands. Group_1948, which encodes for type I CRISPR-associated protein Cas8a1/Csx8, group_2260, which encodes for sodium/solute symporter, and group_2788, which encodes for a hypothetical protein with an unknown function were more prevalent in positive blood culture category. No homologs of group_2260 and group_2788 were detected along genomic islands. Moreover, group_1749, which encodes for POP1 domain-containing protein was detected in all LS-associated genomes and half of BAC-associated ones. However, further investigations are needed to evaluate the exact function of these genes.

Finally, 3 clusters of similar genomic islands were confirmed to be important to identify PBC genomes compared to LI: GI_10, GI_13, GI_45 (Supplementary File 1, Figure S11). However, all of them were located along genomes of Clade A suggesting that the correlation between these clusters of genomic islands and genomes associated with a more severe clinical category might be primarily driven by the clade. Moreover, the Chi-squared test did not support the observation.

Discussion

The exact mechanism by which *F. necrophorum* induces a range of clinical presentations in humans, including the potentially life-threatening systemic disease, is not yet accurately understood. In this study, we searched for genetic determinants that may be correlated with the variable course of the infection and used as biomarkers of

Orthogroup (PBC vs. LI)	Gene name	Gene product	Protein mean length (sd)	Prevalence/PBC genomes	Prevalence/LI genomes	p-value
group_13	–	DUF1353 domain-containing protein	149.5 (0.9)	9/19	40/51	0.026
group_1948	cas8a1	Type I CRISPR-associated protein Cas8a1/Csx8	414.8 (96.5)	15/19	20/51	0.007
group_1955	ftsY	Signal recognition particle-docking protein FtsY	77.1 (5.8)	2/19	26/51	0.005
group_2124	–	T2SP-E domain-containing protein	126	5/19	19/51	0.566
group_2260	–	Sodium:solute symporter	490	8/19	7/51	0.025
group_2788	–	Hypotethical protein	93	4/19	0/51	0.005
Orthogroup (LS vs. BAC)	**Gene name**	**Gene product**	**Mean length (sd)**	**Prevalence/LS genomes**	**Prevalence/BAC genomes**	**p-value**
group_1749	–	POP1 domain-containing protein	407.9 (122.1)	11/11	4/8	0.038

Table 2. Significant orthogroups identified by Boruta in clinical categories comparison. The table displays the relevant orthogroups identified by Boruta to distinguish PBC-associated compared to LI-associated genomes, and LS-associated compared to BAC-associated genomes. For each orthogroup, the gene name, whether available, the gene product, the mean length and standard deviation of the predicted protein, and the prevalence of the orthogroup in the genomes of the clinical categories are reported. The p-values of the Chi-squared test used to evaluate the distribution of orthogroups according to the clinical categories are displayed.

pathogenicity. A few significant associations between genetic determinants and the severity of the disease were observed in our cohort.

Among the orthogroups, a sodium-solute symporter was associated with bacteremia. Its role in cellular homeostasis and nutrient uptake might enhance the adaptive response of the bacteria[56]. Similarly, the CRISPR-associated protein Cas8a1/Csx8 protein was more frequently found in bacteremic strains than in those associated with localized infections. A further characterization of the CRISPR-Cas system types of *F. necrophorum* could reveal a correlation with the clinical presentations. Instead, DUF1353 domain-containing protein and the signal recognition particle-docking protein FtsY were associated with genomes of localized infections. FtsY protein is important for the targeting and translocation of membrane proteins[57], but its role in the virulence of *F. necrophorum* remains unknown. Additionally, the location of these two orthogroups along genomic islands may highlight the contribution of horizontal gene transfer in niche adaptation.

Due to the lack of virulence factor genes specifically associated with *F.necrophorum* in publicly available databases, we included several such databases and further prepared an ad-hoc *F. necrophorum* VF database. We confirmed the ubiquitous presence of known virulence genes of *F. necrophorum subsp. funduliforme* of human origin, such as ecotin, lipopolysaccharide (LPS), and hemolysin; and the absence of FadA. However, no OVFs were associated with the high severity of the infection, but rather with the different clades. The virulome of *F. necrophorum* strains was highly conserved across all the clinical categories, with a higher diversity for the LI genomes. We hypothesize that this could reflect the variability of the infection sites and the consequent need for adaptation to different environments or types of host cells. These results are in accordance with the study of Thapa et al.[58] who reported weak correlation of virulence factors with the source of isolation and lack of phylogenetic clustering of strains from the same source.

The strains analyzed in this study were grouped into two major distinct clades. Clade A contained a significantly higher proportion of bacteremic strains. This phylogenetic structure fitted the phylogeny of *F. necrophorum* previously suggested by Thapa et al. and Bista et al.[58,59]. Indeed, based on 18 publicly available genomes used in all three studies, our clades A and B correspond to Clade B and Clade C, and Clade I and Clade III, respectively in their studies.

The profiles of sequence conservation of many orthogroups, including OVFs, as well as their presence/absence reflected the phylogenetic structure rather than the clinical presentations. Additionally, even the distribution of clusters of genomics islands resembled the phylogeny built on concatenated alignments of single-copy core orthologs. Indeed, we could not identify any pathogenic advantage conferred by the acquisition of genomics islands in our dataset. A higher number of isolates and a reduced number of features might increase the power of GWAs and reveal undiscovered genetic determinants. The clinical isolates provided by Perry et al.[60] would be a precious resource to increase the sample size. However, due to the current lack of patient diagnosis information for the provided strains, this dataset was not used in our study.

The *lktA* gene and the broader *lkt* operon are considered a primary virulence factor recognized to induce apoptosis in leukocytes that should hence favor the spread of the infection. The *lktA* gene encodes for a water-soluble membrane-spanning exotoxin that has an affinity for leukocytes and induces apoptosis of macrophages and neutrophils. LktB is potentially involved in the secretion of LktA due to the presence of a POTRA2 polypeptide transport-associated domain, while LktC has been hypothesized to regulate leukotoxin production due to the presence of histidine kinase and sensory transduction regulator domains[61]. An association of the lktA gene variant type 1b with genomes of bacteremic strains (PBC category) was identified. This variant presented a deletion that generates a premature stop codon around position 3900 and results in the prediction of 2 ORFs along the *lktA* sequence. This deletion has been previously described in *F. necrophorum* subs. *funduliforme* by Holm et al.[53] and was significantly correlated with isolates categorized in the "Head and neck primary focus or diagnosis" group as opposed to the "Non head-and-neck primary focus or diagnosis" group. However, this grouping does not fit our classification since it is primarily based on the localization of the infection, rather than the severity of the infection. Narayanan et al.[52] suggested that the active region of *lktA* may be located within positions 1 and 3696. Thus, the truncated leukotoxin of 3963 bp observed in part of Clade A may achieve a tertiary folding that leaves the active region more exposed, hence increasing its toxicity that would in turn explain the correlation of *lktA* type 1b variant with PBC genomes in our study[62]. Further analyses are needed to understand if the two genes are translated into a protein and whether one of them or together presents a functional folding pattern. However, the limited number of studies aimed at exploring the molecular structure and role of *F. necrophorum* leukotoxin prevents further speculation about the effect of such frameshift in the gene.

Moreover, although differences in the gene sequences of *lkt* operon were already proposed in other studies[53,63,64], we could confirm a strong clade-specific sequence identity[53]. Indeed, the *lktA* type 1b gene variant appeared only in members of Clade A, and exhibited a higher sequence similarity with Clade A variants compared to the representative member of Clade B. This suggests that after the divergence of the two clades, further mutations accumulated in Clade A resulted in the generation of *lktA* type 1b gene variant. Variations in the promoter region and *lkt* genes in the two clades could potentially lead to the misregulation of *lktA* expression, whose increased transcription is associated with high virulence and to a more toxic leukotoxin effect[64,65].

Finally, due to the study design, only strains from patients with a localized or disseminated infection were included. Since strains from fully asymptomatic subjects were omitted, this raises the possibility that the selected strains may inherently possess virulence factors essential for *F. necrophorum* to cause disease. To comprehensively identify these virulence factors, it will be necessary to include strains from asymptomatic subjects in future investigations. Moreover, initial localized infections may of course spread to the bloodstream not only as a consequence of more virulent bacteria, but also according to the timing of effective antibiotic treatment, potentially thus explaining that some strains exhibiting virulence factors may also sometimes be observed among the subgroup of patients with uncomplicated localized infection.

Conclusions

In conclusion, (i) clade A was significantly associated with bacteremia, and (ii) a few genetic determinants, such as the sodium-solute symporter and CRISPR-associated Cas8a1/Csx8 protein, were identified as indicators of pathogenicity in *F. necrophorum* subsp. *funduliforme* across various clinical presentations. However, a larger collection of isolates from multiple sites across the world may better support the identification of additional clades, and confer more statistical power to prove the observed association between *lktA* type 1b variant and bacteremia. The presence of a truncated protein in subjects with bacteremia is unexpected for a predicted virulence factor, but we speculate that the truncation might increase the virulence by exposing specific epitopes. Deletion-associated enhanced virulence is well known for *Shigella*, which is more virulent due to the absence of the *cadA* that encodes the CadA protein known to reduce the effect of the Shiga-enterotoxin[66]. Such a dataset would also allow us to evaluate the impact of other co-factors, among which the timing of infection at the time of diagnosis, host factors (e.g. immunosuppression, host genetics), and possible co-infections. Finally, additional studies are required to understand the function of poorly annotated orthogroups associated with clinical categories and to assess the role of the mutated *lktA* gene, its folding, and its toxicity towards host cells.

Data availability

The sequencing data generated during the current study have been deposited in the European Nucleotide Archive and are available under the BioProject accession number PRJEB74626.

Received: 22 April 2024; Accepted: 19 August 2024
Published online: 27 August 2024

References

1. Langworth, B. F. *Fusobacterium necrophorum*: Its characteristics and role as an animal pathogen. *Bacteriol. Rev.* **41**(2), 373–390 (1977).
2. Shinjo, T., Miyazato, S., & Kiyoyama, H. Adherence of *Fusobacterium necrophorum* biovar a and b strains to erythrocytes and tissue culture cells. In *Annales de l'Institut Pasteur/Microbiologie*, vol. 139, 453–460 (Elsevier, 1988).
3. Riordan, T. Human infection with *Fusobacterium necrophorum* (necrobacillosis), with a focus on Lemierre's syndrome. *Clin. Microbiol. Rev.* **20**(4), 622–659 (2007).
4. Lemierre, A. On certain septicaemias due to anaerobic organisms. *The Lancet* **227**(5874), 701–703 (1936).
5. Wright, W. F., Shiner, C. N. & Ribes, J. A. Lemierre syndrome. *South. Med. J.* **105**(5), 283–288 (2012).
6. Aliyu, S. *et al.* Real-time PCR investigation into the importance of *Fusobacterium necrophorum* as a cause of acute pharyngitis in general practice. *J. Med. Microbiol.* **53**(10), 1029–1035 (2004).
7. Jensen, A., Kristensen, L. H. & Prag, J. Detection of *Fusobacterium necrophorum* subsp. funduliforme in tonsillitis in young adults by real-time PCR. *Clin. Microbiol. Infect.* **13**(7), 695–701 (2007).
8. Windt, D., Kornegoor, R., Walhof, R., Overbeek, B. & Paarlberg, K. Septicaemia with Fu-sobacterium necrophorum from periodontal disease in pregnancy resulting in immature birth: Case report and review of literature. *Obstet. Gynecol. Cases Rev.* **5**, 116 (2018).
9. Brook, I. The role of anaerobic bacteria in tonsillitis. *Int. J. Pediatr. Otorhinolaryngol.* **69**(1), 9–19 (2005).
10. Brazier, J., Hall, V., Yusuf, E. & Duerden, B. *Fusobacterium necrophorum* infections in England and Wales 1990–2000. *J. Med. Microbiol.* **51**(3), 269–272 (2002).
11. Björk, H., Bieber, L., Hedin, K. & Sundqvist, M. Tonsillar colonisation of *Fusobacterium necrophorum* in patients subjected to tonsillectomy. *BMC Infect. Dis.* **15**(1), 1–6 (2015).
12. Allen, B. W., Anjum, F., & Bentley, T. P. Lemierre syndrome. [updated 2023 jul 31]. In Statpearls [internet] (Statpearls Publishing, 2024). https://www.ncbi.nlm.nih.gov/books/NBK499846/
13. Nygren, D. & Holm, K. Invasive infections with *Fusobacterium necrophorum* including Lemierre's syndrome: An 8-year Swedish nationwide retrospective study. *Clin. Microbiol. Infect.* **26**(8), 1089-e7 (2020).
14. Nagaraja, T., Narayanan, S., Stewart, G. & Chengappa, M. *Fusobacterium necrophorum* infections in animals: Pathogenesis and pathogenic mechanisms. *Anaerobe* **11**(4), 239–246 (2005).
15. Narayanan, S. *et al. Fusobacterium necrophorum* leukotoxin induces activation and apoptosis of bovine leukocytes. *Infect. Immun.* **70**(8), 4609–4620 (2002).
16. Umaña, A. *et al.* Utilizing whole fusobacterium genomes to identify, correct, and characterize potential virulence protein families. *J. Bacteriol.* **201**(23), 10–1128 (2019).
17. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120 (2014).
18. Bankevich, A. *et al.* Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**(5), 455–477 (2012).
19. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. Quast: Quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–1075 (2013).
20. Rissman, A. I. *et al.* Reordering contigs of draft genomes using the mauve aligner. *Bioinformatics* **25**(16), 2071–2073 (2009).
21. Schwengers, O. *et al.* Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genomics* **7**(11), 000685 (2021).
22. Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G. & Toth, I. K. Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. *Anal. Methods* **8**(1), 12–24 (2016).
23. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**(6), 1005595 (2017).
24. Marquis, B., Pillonel, T., Carrara, A. & Bertelli, C. zDB: Bacterial comparative genomics made easy. *mSystems* **9**, e00473-24 (2024).
25. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The cog database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**(1), 33–36 (2000).
26. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **42**(D1), 222–230 (2014).
27. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**(7), 1641–1650 (2009).
28. Emms, D. M. & Kelly, S. Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
29. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**(W1), 293–296 (2021).

30. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. Cluster: Cluster analysis basics and extensions (2023). https://CRAN.R-project.org/package=cluster
31. Dudek, A., & Walesiak, M. The choice of variable normalization method in cluster analysis. In *Proceeding of the 35th International Business Information Management Association Conference (IBIMA)* 325–340 (2020).
32. Sayers, S. *et al.* Victors: A web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res.* **47**(D1), 693–700 (2019).
33. Chen, L. *et al.* VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**(suppl–1), 325–328 (2005).
34. Winnenburg, R. *et al.* Phi-base: A new database for pathogen host interactions. *Nucleic Acids Res.* **34**(suppl–1), 459–464 (2006).
35. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**(1), 45–48 (2000).
36. Gillespie, J. J. *et al.* Patric: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **79**(11), 4286–4298 (2011).
37. Camacho, C. *et al.* Blast+: Architecture and applications. *BMC Bioinform.* **10**, 1–9 (2009).
38. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**(4), 772–780 (2013).
39. Robinson, O., Dylus, D. & Dessimoz, C. Phylo. io: Interactive viewing and comparison of large phylogenetic trees on the web. *Mol. Biol. Evol.* **33**(8), 2163–2166 (2016).
40. Bertelli, C. *et al.* Enabling genomic island prediction and comparison in multiple genomes to investigate bacterial evolution and outbreaks. *Microb. Genomics* **8**(5), 000818 (2022).
41. Wishart, D. S., Han, S., Saha, S., Oler, E., Peters, H., Grant, J. R., Stothard, P., & Gautam, V. PHASTEST: Faster than PHASTER, better than PHAST. *Nucleic Acids Res.*, **382** (2023)
42. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**(7), 1575–1584 (2002).
43. Ondov, B. D. *et al.* Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**(1), 1–14 (2016).
44. Verdugo, R., & Orostica, K. chromplot: Global visualization tool of genomic data. *R package version*, **1** (2019)
45. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant 868 bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, 15 (2015).
46. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. Pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**(24), 4310–4312 (2018).
47. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: Snps in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**(2), 80–92 (2012).
48. Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.***3** (2012)
49. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
50. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/ (2013)
51. Erguler, K. Barnard: Barnard's Unconditional Test. R package version 1.8.1. https://github.com/kerguler/barnard (2023).
52. Narayanan, S. K., Nagaraja, T., Chengappa, M. & Stewart, G. C. Cloning, sequencing, and expression of the leukotoxin gene from *Fusobacterium necrophorum*. *Infect. Immun.* **69**(9), 5447–5455 (2001).
53. Holm, K., Collin, M., Hagelskjær-Kristensen, L., Jensen, A. & Rasmussen, M. Three variants of the leukotoxin gene in human isolates of *Fusobacterium necrophorum* subspecies funduliforme. *Anaerobe* **45**, 129–132 (2017).
54. Hacker, J. & Carniel, E. Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO Rep.* **2**(5), 376–381 (2001).
55. Susko, E. & Roger, A. J. Long branch attraction biases in phylogenetics. *Syst. Biol.* **70**(4), 838–843 (2021).
56. Henriquez, T., Wirtz, L., Su, D. & Jung, H. Prokaryotic solute/sodium symporters: Versatile functions and mechanisms of a transporter family. *Int. J. Mol. Sci.* **22**(4), 1880 (2021).
57. Seluanov, A. & Bibi, E. Ftsy, the prokaryotic signal recognition particle receptor homologue, is essential for biogenesis of membrane proteins. *J. Biol. Chem.* **272**(4), 2053–2055 (1997).
58. Thapa, G. *et al.* A genome-led study on the pathogenesis of *Fusobacterium necrophorum* infections. *Gene* **840**, 146770 (2022).
59. Bista, P. K., Pillai, D., Roy, C., Scaria, J. & Narayanan, S. K. Comparative genomic analysis of *Fusobacterium necrophorum* provides insights into conserved virulence genes. *Microbiol. Spectr.* **10**(6), e00297-22 (2022).
60. Perry, M. D. *et al.* First large-scale study of antimicrobial susceptibility data, and genetic resistance determinants, in *Fusobacterium necrophorum* highlighting the importance of continuing focused susceptibility trend surveillance. *Anaerobe* **80**, 102717 (2023).
61. Tadepalli, S., Stewart, G. C., Nagaraja, T. & Narayanan, S. K. Leukotoxin operon and differential expressions of the leukotoxin gene in bovine *Fusobacterium necrophorum* subspecies. *Anaerobe* **14**(1), 13–18 (2008).
62. Wright, K. Genomics and virulence factors of *Fusobacterium necrophorum*. https://api.semanticscholar.org/CorpusID:90029571 (2016).
63. Zhang, F., Nagaraja, T., George, D. & Stewart, G. C. The two major subspecies of *Fusobacterium necrophorum* have distinct leukotoxin operon promoter regions. *Vet. Microbiol.* **112**(1), 73–78 (2006).
64. Tadepalli, S., Stewart, G. C., Nagaraja, T. & Narayanan, S. K. Human *Fusobacterium necrophorum* strains have a leukotoxin gene and exhibit leukotoxic activity. *J. Med. Microbiol.* **57**(2), 225–231 (2008).
65. Pillai, D. K., Amachawadi, R. G., Baca, G., Narayanan, S. K. & Nagaraja, T. Leukotoxin production by *Fusobacterium necrophorum* strains in relation to severity of liver abscesses in cattle. *Anaerobe* **69**, 102344 (2021).
66. Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. "Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of shigella spp. and enteroinvasive *Escherichia coli*. *Proc. Natl. Acad. Sci.* **95**(7), 3943–3948 (1998).

Acknowledgements

We would like to thank Maria Senra-Ortiz, Anne-Laure Chanson, and Rizlène Dira for their technical help.

Author contributions

CB, CG, JS, TP, and GG conceived the project. AC analyzed the Fusobacterium genomes with the help of TP and under the co-supervision of CB, TP, and GG. Clinical metadata were retrieved by ADO and JS for patients hospitalized in Geneva and by CG, GG, and BM for patients from Lausanne. Statistical analyses were mainly performed by AC with co-supervision of CB, TP, and GG. AC wrote a first draft of the manuscript which was corrected and approved by all co-authors.

Funding
This project has been mainly supported by a dedicated budget for applied research in diagnostic microbiology (IMUR 28806), which aimed at developing further bacterial genomics and especially the tools to detect virulence factors in clinically relevant infections. The salary of Alessia Carrara has been supported by Greub's research funding (IMUR 31982 & 28806) and C. Bertelli's research funding. In addition, the salary of Bastian Marquis (MD PhD student in Greub's group) was funded by the Jürg Tschopp MD-PhD scholarship.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70608-y.

Correspondence and requests for materials should be addressed to G.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.