communications biology

Article

https://doi.org/10.1038/s42003-024-06825-y

Comparative genomics reveal a novel phylotaxonomic order in the genus *Fusobacterium*

Check for updates

Cristian Molteni ✉, Diego Forni, Rachele Cagliani & Manuela Sironi

Fusobacteria have been associated to different diseases, including colorectal cancer (CRC), but knowledge of which taxonomic groups contribute to specific conditions is incomplete. We analyzed the genetic diversity and relationships within the *Fusobacterium* genus. We report recent and ancestral recombination in core genes, indicating that fusobacteria have mosaic genomes and emphasizing that taxonomic demarcation should not rely on single genes/gene regions. Across databases, we found ample evidence of species miss-classification and of undescribed species, which are both expected to complicate disease association. By focusing on a lineage that includes *F. periodonticum/pseudoperiodonticum* and *F. nucleatum*, we show that genomes belong to four modern populations, but most known species/subspecies emerged from individual ancestral populations. Of these, the *F. periodonticum/pseudoperiodonticum* population experienced the lowest drift and displays the highest genetic diversity, in line with the less specialized distribution of these bacteria in oral sites. A highly drifted ancestral population instead contributed genetic ancestry to a new species, which includes genomes classified within the *F. nucleatum animalis* diversity in a recent CRC study. Thus, evidence herein calls for a re-analysis of *F. nucleatum animalis* features associated to CRC. More generally, our data inform future molecular profiling approaches to investigate the epidemiology of *Fusobacterium*-associated diseases.

Fusobacteria are Gram-negative, non-spore-forming obligate anaerobes with a wide distribution. The phylum Fusobacteriota includes both species commonly found in animal microbiota and others that are free-living in the marine environment[1]. Within the phylum, members of the genus *Fusobacterium* are found in the mouth and other mucosal sites of humans and other animals[1]. In the human oral cavity, *Fusobacterium* species participate to the formation of polymicrobial biofilms and are associated with periodontal disease. Remarkably, these bacteria have the ability to spread to extraoral sites where they contribute to the development of different conditions, including Lemierre syndrome, appendicitis, brain abscesses, osteomyelitis, pericarditis, inflammatory bowel disease, and cancer[1].

Fusobacterium species have gained enormous interest in relation to their potentially pathogenic role in colorectal cancer (CRC) and other tumor types. In particular, most studies have focused on *Fusobacterium nucleatum*, which was shown to be enriched in the gut microbiota of CRC patients and to promote carcinogenesis through multiple mechanisms[1–6]. *F. nucleatum* (and *F. necrophorum)* was also found in the lymph node and liver metastases of *Fusobacterium*-associated primary tumors[3]. In addition to its role in CRC, *F. nucleatum* has been involved in other cancer types (bladder, oral, head and neck, cervical, and gastric) and in periodontitis[1,7]. Also, *F. nucleatum* can colonize the placenta and cause preterm birth, intra-amniotic infection, stillbirth, neonatal sepsis, and hypertensive disorders of pregnancy[1,8,9]. Growing evidence however suggests that *Fusobacterium* species other than *F. nucleatum* associate with CRC[3,10–13]. Moreover, *F. nucleatum* bacteria are presently classified into four subspecies (*nucleatum*, *animalis*, *vincentii*, and *polymorphum)*. These subspecies are phylogenetically divergent to the point that they were suggested to represent distinct species[14,15]. Also, a recent report suggested that *F. nucleatum* subspecies *animalis* is divided into two clades, only one of which is associated with CRC[6]. Finally, several works identified species miss-classifications in public records, whereas some *Fusobacterium* genomes cannot be assigned to any existing species[12,16].

Recently, it was suggested that phylogenetic analyses based on the *rpoB* gene, rather than on 16 s rRNA, are better suited to differentiate *Fusobacterium* species and to classify genomes into lineages[12]. However, analyses in several human commensal microbiota and environmental bacteria have suggested that homologous recombination may affect the majority of loci in the genome[17–23]. Thus, for many species, each locus has recombined extensively, and consequently the phylogeny changes many times along the

Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy. ✉e-mail: cristian.molteni@lanostrafamiglia.it

genome alignment, making it impossible to reconstruct robust clonal relationships. This is also the case of fusobacteria, as different studies documented horizontal gene transfer (HGT) and recombination[24–26]. Nonetheless, the extent of recombination in the extended *Fusobacterium* genus has not been investigated, making it difficult to assess how well single gene-based phylogenies can represent the relationships among genomes. Also, a comprehensive analysis of the genetic diversity and of evolutionary relationships in this genus is presently missing.

Results

Recombination in rpoB and relevance for lineage definition

The *rpoB* gene was recently suggested to represent a good marker for the classification and the phylogenetic reconstruction of relationships among members of the genus *Fusobacterium*[12]. We thus retrieved from public databases sequence information for 361 *Fusobacterium* genomes and we extracted *rpoB* sequences, which were identified for 345 strains (see "Methods" section). The neighbor-net split network of *rpoB* showed a complex reticulation pattern, suggesting extensive recombination (Fig. 1). In line with previous reports, *F. naviforme* sequences, as well as some other unassigned species, were highly divergent[10,12,27]. Given the observed reticulation, we used the fastGEAR software to identify and analyze recombination events. This software first classifies the sequences into lineages; subsequently, it calculates the number of ancestral and recent recombination events and tests for their significance. fastGEAR divided *rpoB* sequences into 12 lineages and identified 198 ancestral recombination events and 98 recent ones (Fig. 1, Supplementary Fig. 1, and Supplementary Data 1). This clearly indicates the presence of extensive recombination and implies that different gene regions have distinct evolutionary histories. We thus used SimPlot analysis to generate a sequence similarity network based on *rpoB* sequences, which joins nodes (sequences or groups of sequences) with edges when similarity is above a given threshold. With a threshold for global and local similarities at 95%, the twelve lineages remained separated (Fig. 2A). However, several regions of local similarity above 95% were detected, in line with the effects of recombination.

On one hand, the lineage subdivision generated by fastGEAR showed good agreement with the clusters in the neighbor-net split network, with the exclusion of lineage 7, which was split into two clusters (Fig. 1). On the other, the identified lineages only partially reflected the demarcation of known *Fusobacterium* species and two lineages (1 and 4) were only populated with unclassified fusobacteria. In several instances, more than one species was classified in the same lineage, whereas in the case of *F. varium* and *F. perfoetens*, genomes were split into two different lineages (Fig. 1). Specifically, one *F. varium* sequence (strain An876) was assigned to lineage 7 together with some unassigned species and with *F. hominis*, while all the others were in lineage 9, which also includes *F. ulcerans*. Likewise, one *F. perfoetens rpoB* sequence (strain An877) and several unassigned species were in lineage 8, whereas the remaining ones were assigned to lineage 6. Finally, multiple species were detected in lineages 3 (*F. equinum* and *F. gonidiaformans*) and 5 (*F. russii, F. massiliense, F. gastrosuis*) (Fig. 1).

We thus used SimPlot to analyze the similarity between the *rpoB* sequences in these lineages. Briefly, results (Fig. 2B) indicated that (i) the *F. varium* sequence in lineage 7 shows less than 95% local and global similarity to other *F. varium* sequences in lineage 9 and the same holds true for the *F. perfoetens* sequence in lineage 8 compared to other *F. perfoetens* sequences; thus, these two sequences are likely to be misclassified; (ii) *F. equinum* and *F. gonidiaformans* have high sequence similarity (>95%) (see below); (iii) *F. russii, F. massiliense*, and *F. gastrosuis* display below threshold similarity at the global and local level. Overall, results based on *rpoB* sequences confirm previous indications that the *Fusobacterium* genus includes substantial unclassified diversity and that some sequences are miss-classified.

Finally, we compared the classification determined by fastGEAR with the nine lineages defined by Bi and coworkers using *rpoB* sequences[12].

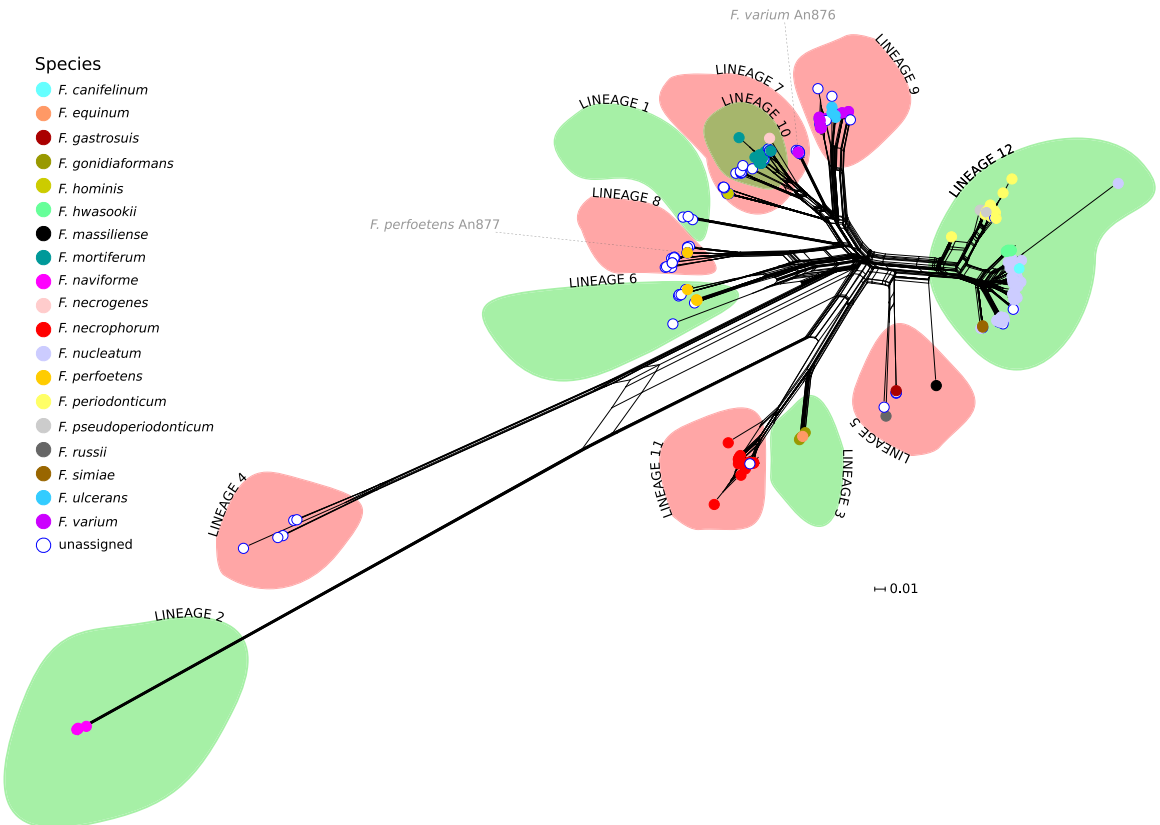

Fig. 1 | Recombination and divergence among *Fusobacterium* species. Neighbor-net split network of 345 *rpoB* genes. Each sequence is shown as a dot, color-coded by species. The green and red areas represent the lineages defined by fastGEAR analysis (see "Methods" section). The *F. varium* An876 and *F. perfoetens* An877 sequences are highlighted in gray (see also Figs. 2B and 4).

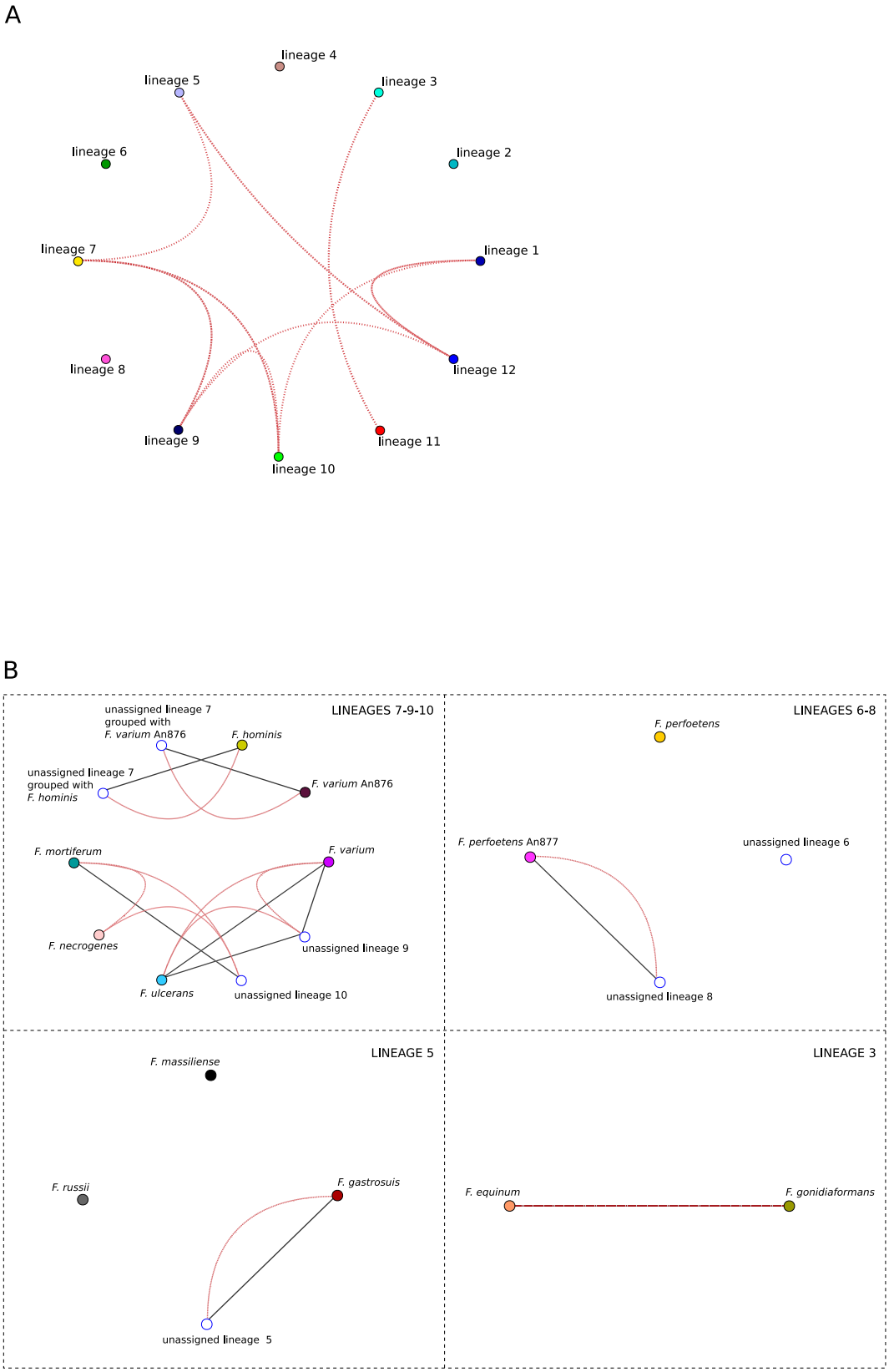

Fig. 2 | Sequence similarities among lineages. A Sequence similarity network based on *rpoB* genes. Sequences are grouped by lineages, as defined by fastGEAR. Global sequence similarity is represented by black edges and it is not observed among *rpoB* sequences. Red edges represent local similarity. Thresholds for global and local similarities were set to 95%. **B** Sequence similarity networks based on *rpoB* sequences are shown for selected lineages (see main text). As in panel **A**, lineages are color-coded as in Fig. 1. Thresholds for global and local similarities were set to 95%.

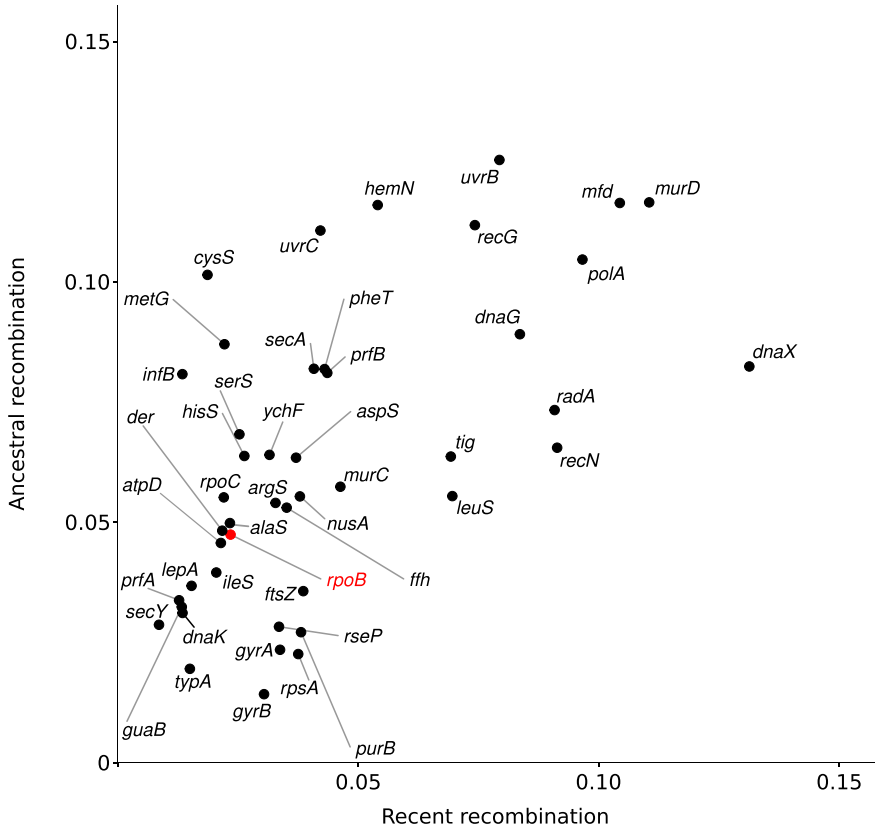

Fig. 3 | Recombination intensities in core genes. Correlation between ancestral and recent recombination events for 45 core genes. Recombination events were calculated with fastGEAR and divided by alignment length. Each dot is labeled with the corresponding gene name and *rpoB* is highlight in red.

This exercise was complicated by the fact that the analyzed sequences differ. Nonetheless, a relatively good correspondence was found. The main differences related to *F. necrophorum* and *F. gonidiaformans*, which fastGEAR classified in two distinct lineages, whereas they both contributed to the same lineage in Bi et al. On the contrary, fastGEAR classified *F. massiliense* and *F. russii* together whereas they accounted for lineages 2 and 3 in the previous classification proposed by Bi and coworkers[12].

Recombination in core genes and its effects on classification

We next aimed to investigate and compare the levels of recent and ancestral recombination among different core genes. We thus used the Genome Taxonomy Database Toolkit to extract the sequences of 120 core genes present in more than 300 fusobacterial genomes. Among these we retained only the ones longer than 1000 bp ($n = 45$) and used fastGEAR to detect recombination. For these 45 genes, the inferred number of lineages varied from 8 to 19. In most cases, the levels of ancestral recombination were much higher than the recent. The number of ancestral events ranged from 37 to 446, whereas recent events ranged from 12 to 400 (Supplementary Table 1). Consistent with a relatively constant rate of recombination at individual genes, the number of ancient and recent events was highly correlated (Pearson's correlation coefficient= 0.58, p value = 2.4×10^{-05}). Because the amount of recombination events is clearly also a function of gene size, we normalized the number of events by alignment length, so as to have a measure of recombination intensity. The results indicated that *rpoB* is in the low range of recombination intensities, whereas several of the top recombining genes are involved in DNA replication and repair (i.e., *polA*, *mfd, dnaG, recG, uvrB, dnaX, radA, and recN*), as well as in peptidoglycan biosynthesis (*murD*) (Fig. 3) (Supplementary Table 1).

Overall, these results suggest that, in fusobacteria, individual genes have different evolutionary histories and recombination intensities. As a consequence, a classification based on individual genes is expected to be highly sensitive to the choice of the genomic region.

To gain further insight into the effect of gene choice, we again resolved to SimPlot analysis. In particular, we generated a concatenated alignment with the 120 core genes and we analyzed global similarity among species and lineages (Fig. 4). Using a threshold of 95%, most edges joined nodes belonging to lineage 12 (*F. nucleatum*, *F. hwasookii*, *F. canifelinum* and *F. simiae*, as well as *F. periodonticum* and *F. pseudoperiodonticum*) and unassigned sequences therein. In line with a very recent report, *F. equinum* and *F. gonidiaformans* showed high similarity in the extended set of core genes, as well[27] (Fig. 1, Fig. 4). We next compared global with local similarity patterns defined by three genes: *rpoB*, *typA* (with low recombination intensity) and *murD* (with high recombination intensity). In all cases local similarities joined more lineages/species than global similarities. This indicates that classifications based on single genes (i.e., based on local similarities) tend to cluster together sequences that are divergent at the level of the extended set of core genes (i.e., at the level of global similarity) (Fig. 4). Also, whereas the pattern of local similarity was relatively similar for the low recombining *rpoB* and *typA*, it was not for the highly recombining *murD* gene. Indeed, fewer cases of local high similarity were detected with *murD* and in some instances *murD* sequences were more divergent than 95% even between species that showed high global similarity (e.g., *F. simiae* and *F. nucleatum* or *F. hwasookii* and *F. canifelinum*) (Fig. 4).

Finally, we aimed to assess whether phylogenetic reconstruction is affected by gene choice and by the variable level of recombination in core genes. We thus used the Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) program to construct phylogenetic trees that account for the effect of recombination[28]. Specifically, we generated a tree using the *rpoB* alignment and another using the concatenated alignment of core genes. When we used a tanglegram representation to compare the two trees, several entanglements were evident, although most occurred for tips within individual lineages (Fig. 5). Thus, the overall lineage definition obtained with the neighbor-net split network was recapitulated by both trees. Overall, these data underscore the effect of recombination on similarity scores and on the phylogenetic reconstruction of closely related

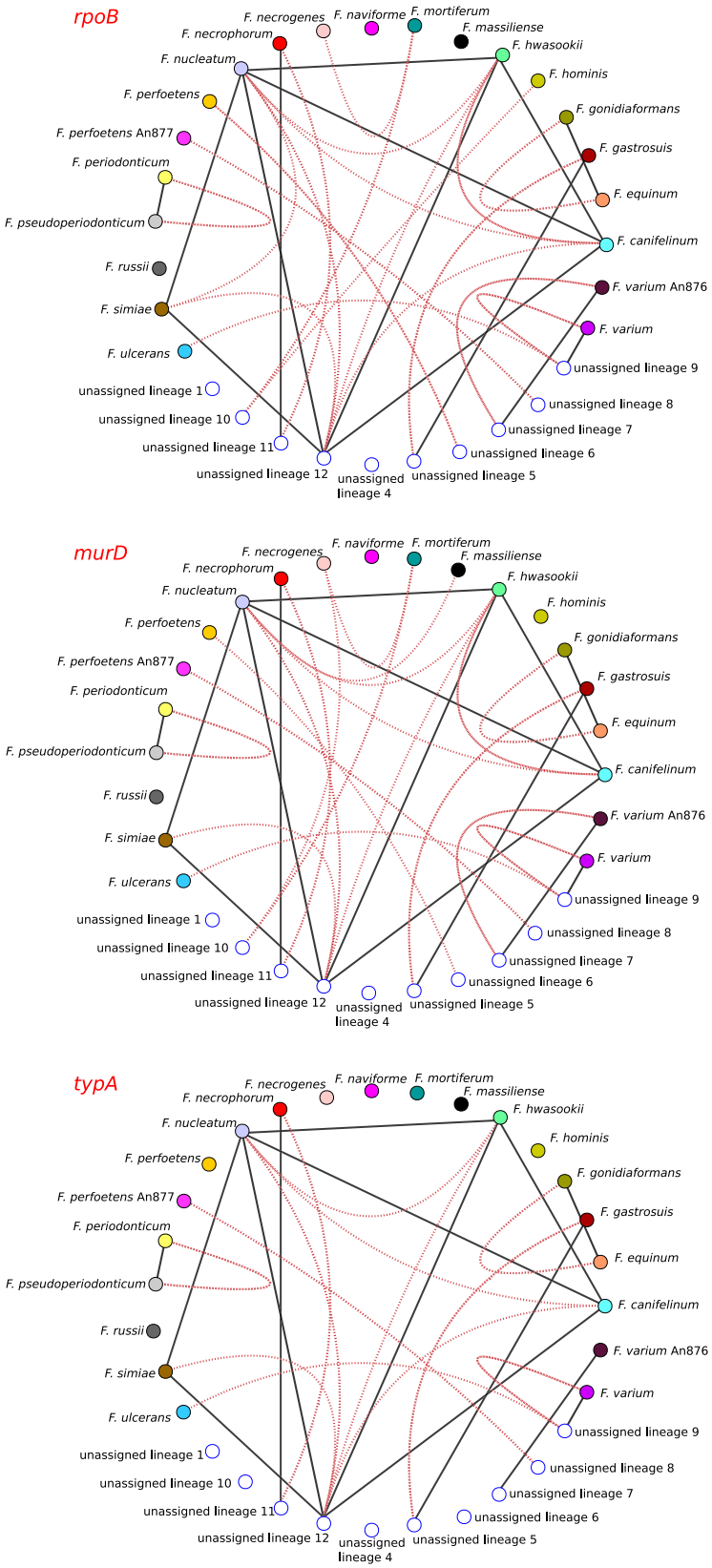

Fig. 4 | Sequence similarities among *Fusobacterium* species and lineages. Sequence similarity networks based on a concatenated core gene alignment. For all networks, global and local similarity thresholds were set to 95%. Black edges represent global sequence similarity (i.e. calculated for the whole concatenated gene alignment). Red edges display local similarity within three different selected genes in the alignment: *rpoB*, *murD*, and *typA*. Each species or lineage is shown as a colored node (colors as in Fig. 1).

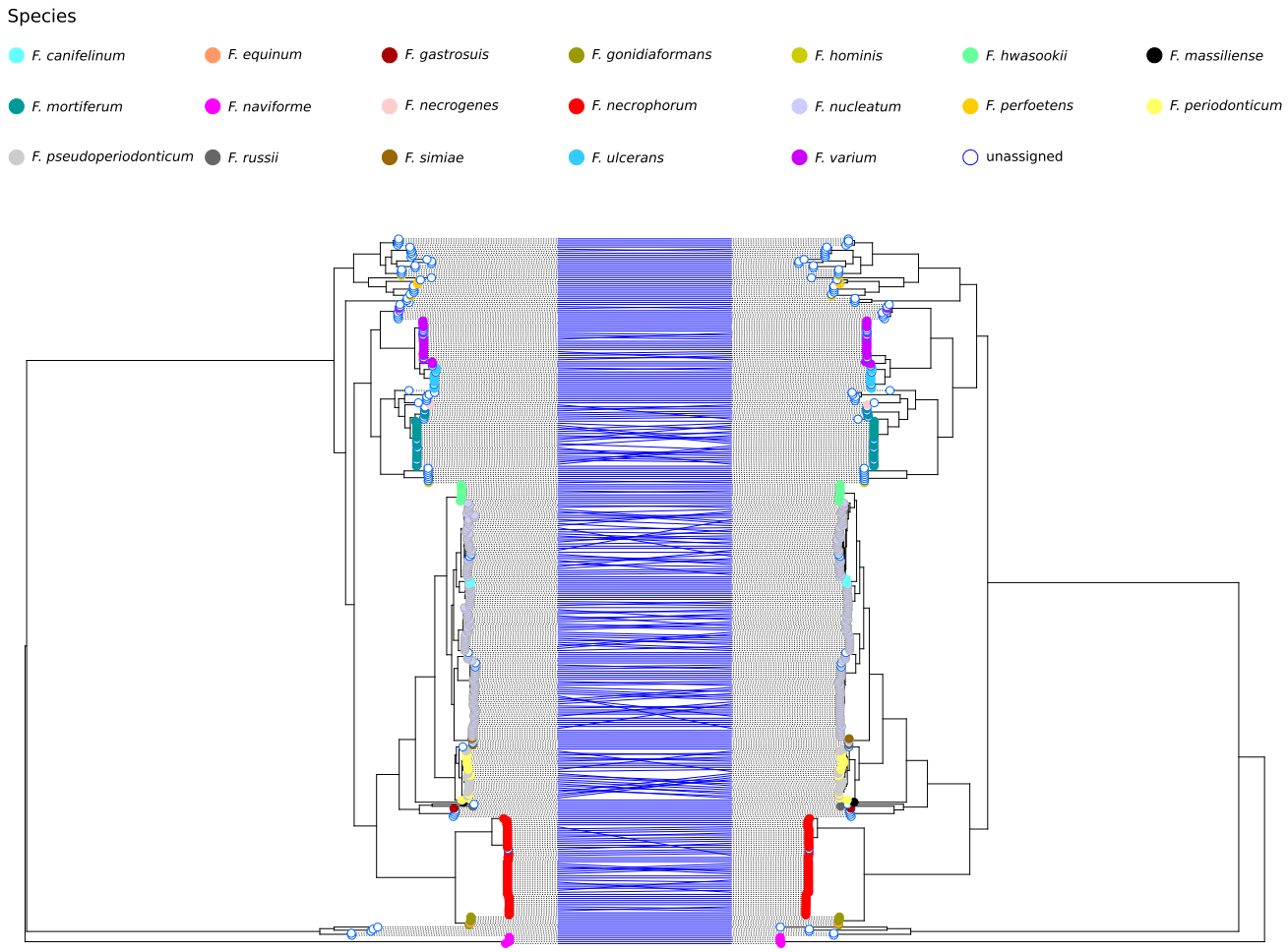

Fig. 5 | Phylogenetic relationships among Fusobacteria species. Cophylogenetic representation of recombination-free phylogenetic trees of 345 Fusobacterium strains. Trees were built on *rpoB* (left) or core genome (right) alignments. Blue lines link the same strain on both trees, tip colors are shown in the legend.

sequences, thus cautioning against the use of individual genes or gene regions for classification purposes.

Genetic relationships in the *F. nucleatum*/*F. periodonticum* lineage

We next aimed to investigate the genetic relatedness of core genomes among fusobacteria in lineage 12 (lineage 1 in Bi and co-workers[12]) (Fig. 1). This lineage comprises species often associated with CRC, including the highly studied *F. nucleatum* (and its subspecies), and the different species have high sequence similarity in the analysis of core genes (Fig. 4). Specifically, these species have an average identity of 92% calculated on the concatenated alignment. We thus extracted parsimony-informative (PI) sites from the core gene alignment. PI information was used as the input for principal component analysis (PCA). In agreement with the neighbor-net split tree, the first PC explained 32% of the variance and separated the two main sub-lineages – i.e., *F. periodonticum/pseudoperiodonticum* from the other *Fusobacterium* species (Fig. 6A). Along this component, *F. periodonticum* separated into two sub-clusters, suggesting the presence of unrecognized diversity within this species. The second PC explained 19% of variance and separated the non-*F. periodonticum/pseudoperiodonticum* core genomes in three provisional clusters: (i) one containing *F. nucleatum polymorphum, F. hwasookii, F. canifelinum* and *F. nucleatum nucleatum*, plus several unclassified strains; (ii) another comprising *F. nucleatum vincentii, F. simiae* and four unclassified genomes; (iii) the third only featuring *F. nucleatum animalis* and unclassified species (Fig. 6A). Analysis of the third PC, which explained 10% of the variance, only revealed some separation of *F. nucleatum vincentii* from the other sequences (Supplementary Fig. 2). This indicates that the four *F. nucleatum* subspecies are less closely related to each other than they are to other *Fusobacterium* species, suggesting that they should be considered as separate species.

The results of the PCA were used to provisionally assign unclassified genomes to known species (Fig. 6A). The only exception was accounted for by four unclassified genomes closely related to each other, suggesting they represent an undescribed species. Indeed, one of these is *Fusobacterium* FNU, previously suggested to represent a new species[16]. Another genome in this hypothetical new species belongs to strain 13-08-02 (BHYR00000000). Very recently, Zepeda-Rivera and coworkers reported that *F. nucleatum animalis* genomes can be divided into two clades referred to as C1 and C2, with the latter associated with the CRC niche[6]. Strain 13-08-02 was included in C1, whereas their clade C2 comprised a number of genomes classified as *F. nucleatum animalis* in NCBI and in the PCA analysis (Fig. 6A). Overall, the PCA does not support the idea that genomes in clade C1 belong to *F. nucleatum* subspecies *animalis*. Indeed, SimPlot analysis confirmed the four clusters identified in the PCA and showed that the hypothetical new species/clade C1 displays 95% similarity to both *F. nucleatum animalis* and to *F. nucleatum vincentii* (Fig. 6B). Most likely, the 95% identity between the new species/clade C1 and *F. nucleatum animalis* (clade C2) is higher than that calculated by Zepeda-Rivera and coworkers (92-93%) because we used only core genes. SimPlot analysis also confirmed the designation of taxonomic levels as species or subspecies to be problematic[14,15]. In fact, the core genomes of some species (e.g., *F. hwasookii* and *F. nucleatum polymorphum* or *F. nucleatum nucleatum* and *F. canifelinum*) were more closely related to each other than subspecies are among themselves. This was also confirmed by a phylogenetic tree generated with Gubbins, that separated all species,

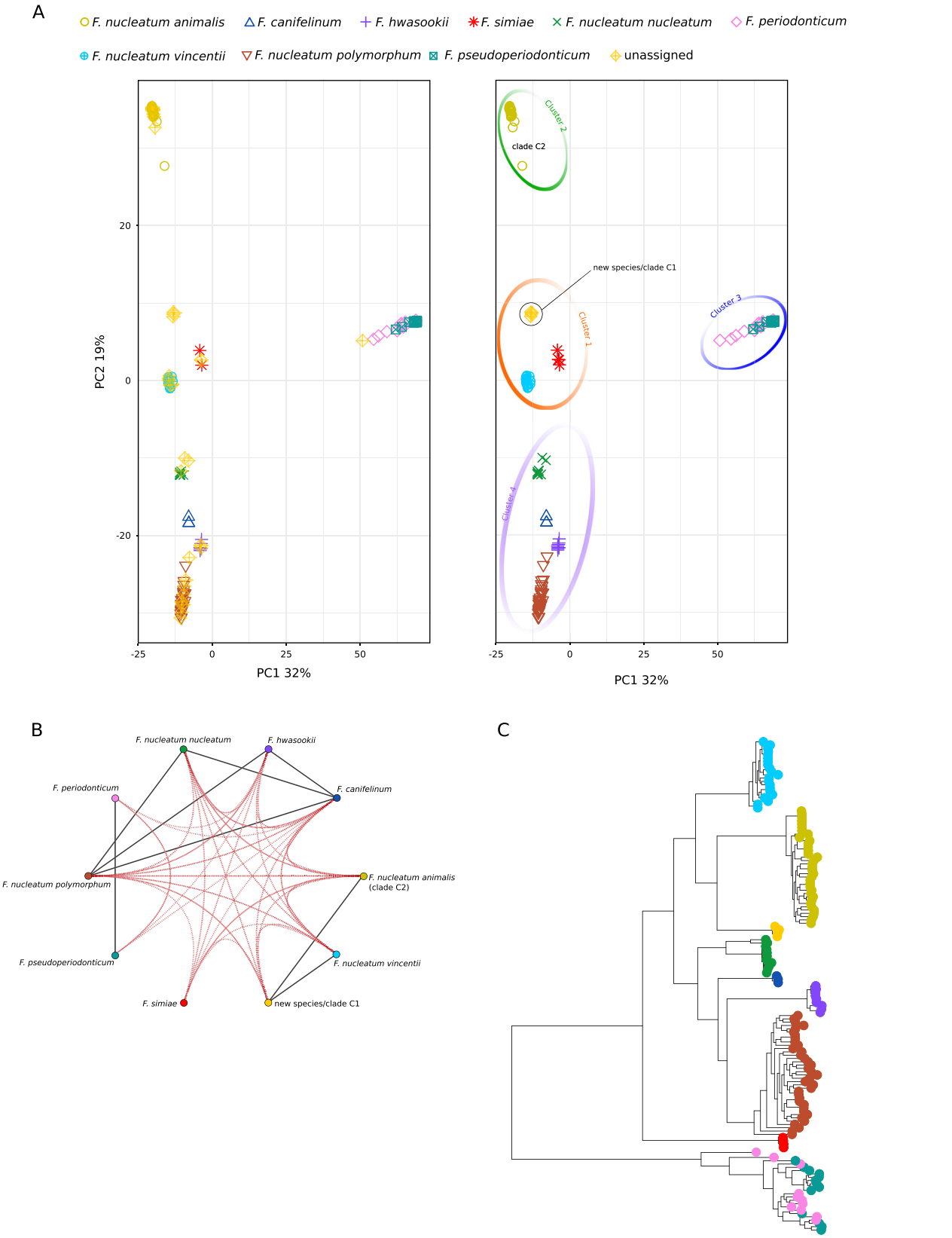

Fig. 6 | Principal Component Analysis (PCA), similarity network, and phylogenetic tree for *Fusobacterium* strains in the *F. nucleatum/F. periodonticum* lineage. A Each *Fusobacterium* genome is colored and displayed with a different symbol, as described in the legend. On the left side, the plot shows several unassigned sequences, which were reassigned in the right panel. In the right panel, the four major clusters are highlighted by colored circles. The genomes contributing to the new species/clade C1 and clade C2 are indicated. **B** Sequence similarity network. Global and local sequence similarity are shown in black and red edges (thresholds were set to 95%). Nodes are colored as in panel **A**. **C** Approximately-maximum-likelihood phylogenetic tree of recombination-free core genes. Tips are colored as in panel **A**.

subspecies and clades (Fig. 6C). The only exception was accounted for by *F. periodonticum* and *F. pseudoperiodonticum* sequences which, as in the PCA, grouped together.

Analysis of accessory genes and complete genomes of the F. nucleatum/F. periodonticum lineage

We next sought to investigate how the genetic relationships established using the core genomes related to accessory gene content and full genome information. We thus used the PATO toolkit[29] to extract accessory genes from genomes in lineage 12. A total of 23,662 accessory genes were obtained, which were used for a principal component analysis. Results indicated a somehow different picture than the one obtained using core genome sequences, with the unclassified species/clade C1 and all *F. nucleatum* subspecies except *F. nucleatum polymorphum* clustering together (Fig. 7A). The PCA also showed that *F. periodonticum* and *F. pseudoperiodonticum* have similar accessory gene contents and the same applies to *F. hwasookii* and *F. nucleatum polymorphum*, with the latter showing some heterogeneity (Fig. 7A). It should however be noticed that the PCA had overall limited discriminatory ability and the first two PC explained only 12% and 9% of the variance (Fig. 7A).

We next moved to the analysis of complete genomes by calculating average nucleotide identities (ANI). A 96% cutoff is often used to define bacterial species using ANI analysis[30]. Using this criterion, we identified 8 species corresponding to *F. nucleatum animalis* (clade C2), *F. nucleatum nucleatum*, the new species/clade C1, *F. nucleatum vincentii, F. simiae, F. nucleatum polymorphum, F. canifelinum*, and *F. hwasookii* (Fig. 7B). Within these species, *F. nucleatum polymorphum* showed the largest heterogeneity. More complex was the situation for bacteria presently classified as *F. periodonticum* and *F. pseudoperiodonticum:* most of them showed identity > 96%, however some were borderline or below threshold and there was no clear separation of genomes classified in the two species (Fig. 7B). Overall, these results indicate that different classifications are obtained using the core genome, the accessory gene content, or ANI calculation over complete genomes.

Population structure of the F. nucleatum/F. periodonticum lineage

To gain further insight into the population structure of fusobacteria core genomes, we used the program STRUCTURE, which relies on a Bayesian statistical model for clustering genotypes into populations, without prior information on their genetic relatedness[31–33]. The program can identify distinct subpopulations (or clusters, K) that compose the overall population. Subpopulations can then be related to specific features such as origin, classification, or phenotype. STRUCTURE is ideally suited to study highly recombining populations[31,34].

Initially, we used the no admixture model, in which each individual is assumed to have derived from one of the modern populations. To estimate the optimal number of subpopulations in the *Fusobacterium* dataset, STRUCTURE was run for values of K from 1 to 12. The ΔK method yielded two peaks at $K = 2$ and $K = 4$ (Supplementary Fig. 3). At $K = 2$, STRUCTURE clearly separated the two main sub-lineages (*F. periodonticum/pseudoperiodonticum* and *F. nucleatum* plus related species) (Fig. 8). At $K = 4$, the four subpopulations paralleled the clusters identified in the PCA, and the new species/clade C1 was assigned to the population that includes *F. nucleatum vincentii* and *F. simiae*, not *F. nucleatum animalis* (Fig. 8).

In order to gain further insight into the evolutionary history of *Fusobacterium* core genomes, we repeated STRUCTURE analysis using the linkage model with correlated allele frequencies. This model assumes that discrete genome "chunks" were inherited from K ancestral populations[32]. The ΔK method identified two major peaks at $K = 3$ and $K = 9$ (Supplementary Fig. 4). We thus analyzed in detail the results at $K = 9$, which represents the finest level of structure for these genomes. Analysis of ancestry components showed that one of the ancestral populations (P_common) contributed variable proportions of ancestry to most genomes (Fig. 9A). Other than this, individual ancestral components accounted most of the ancestry of distinct species or subspecies, with the only exception of *F. canifelinum*, which received ancestry components from 4 populations. Genomes of the new hypothetical species/clade C1 had most of their ancestry accounted by one of the nine ancestral populations, confirming they represent an entity distinct from *F. nucleatum animalis* (Fig. 9A).

The linkage model also allows estimation of the F parameter, which represents a measure of genetic differentiation between populations based on allele frequencies. F can be interpreted as a measure of drift from a hypothetical common ancestral population. The lowest drift was detected for P_common, which is shared among most genomes (Fig. 9B). However, the second population showing lowest drift was the one accounting for most ancestry of *F. periodonticum* and *F. pseudoperiodonticum*. The highest drift was instead obtained for the populations contributing ancestry to *F. simiae* and to the hypothetical new species/clade C1 (Fig. 9B).

We next calculated nucleotide diversity and Tajima's D for core genomes that acquired a major part of their ancestry (>80%) from a single population. In line with the F results, the highest diversity was observed for the *F. periodonticum/F. pseudoperiodonticum* population, which also displayed the most negative value of Tajima's D (Fig. 9C). Overall, this may be suggestive of a genetically diverse population that has expanded in size. Conversely, low diversity was observed for *F. simiae* and the hypothetical new species/clade C1, which both showed moderately negative Tajima's D, possibly suggesting that these populations have expanded after a bottleneck (Fig. 9C). It should however be added that *F. simiae* and the hypothetical new species/clade C1 were represented by very few genomes, raising concerns of representativeness. These results will thus need replication when additional genomic data become available. We should also mention that we cannot exclude that some cryptic population structure in the *F. periodonticum/F. pseudoperiodonticum* population has remained undetected and this might have inflated nucleotide diversity measures and biased the Tajima's D value.

Finally, we exploited the expanded Human Oral Microbiome Database (eHOMD) to compare the distributions of *Fusobacterium* species among sites in the oral cavity and pharynx[35]. Because these data were mostly derived from the typing of two regions (V1-V3 and V3-V5) of the 16S rRNA[36], they are not well suited to distinguish closely related fusobacteria species[12]. We thus considered the four *F. nucleatum* subspecies as a single group to which we added the closely related *F. hwasookii*. Data indicated that *F. nucleatum/hwasookii* bacteria are mostly specialized for the gingival plaque niche, with lower abundance in other sites (Fig. 9D). Conversely, *F. periodonticum* seems to occupy different sites, including hard palate, tongue dorsum, palatine tonsils, throat, and saliva (Fig. 9D). These results may be consistent with the idea that the *F. periodonticum/pseudoperiodonticum* population is more generalist in terms of distribution, whereas the other species/subspecies might have drifted from the ancestral population as a consequence of niche adaptation.

Discussion

Because fusobacteria potentially contribute a huge health burden to human populations, molecular profiling approaches are essential to understand the epidemiology of fusobacteria-associated disease and to relate taxonomic groups to specific conditions. Even in the case of CRC, uncertainty still exists about which *Fusobacterium* species are pathogenic. A recent study showed that a lineage that includes *F. nucleatum* (all subspecies), *F. hwasookii*, *F. periodonticum*, and related fusobacteria is enriched in tumor samples and feces from CRC patients. A different lineage (represented by *F. varium* and *F. ulcerans)* was instead associated with lymphovascular invasion[12]. Conversely, another study found enrichment of *F. varium* in CRC samples[11], whereas *F. nucleatum animalis* or even a specific clade within the diversity of this subspecies was found to be overabundant in the CRC niche in different studies[5,6,37]. Compared to CRC, other fusobacteria-related diseases have been investigated in shallower details and most commonly used approaches do not allow fine taxonomic definition. Moreover, as previously reported, we found that a number of *Fusobacterium* sequences in public repositories are miss-classified and several undescribed species exist, which complicates

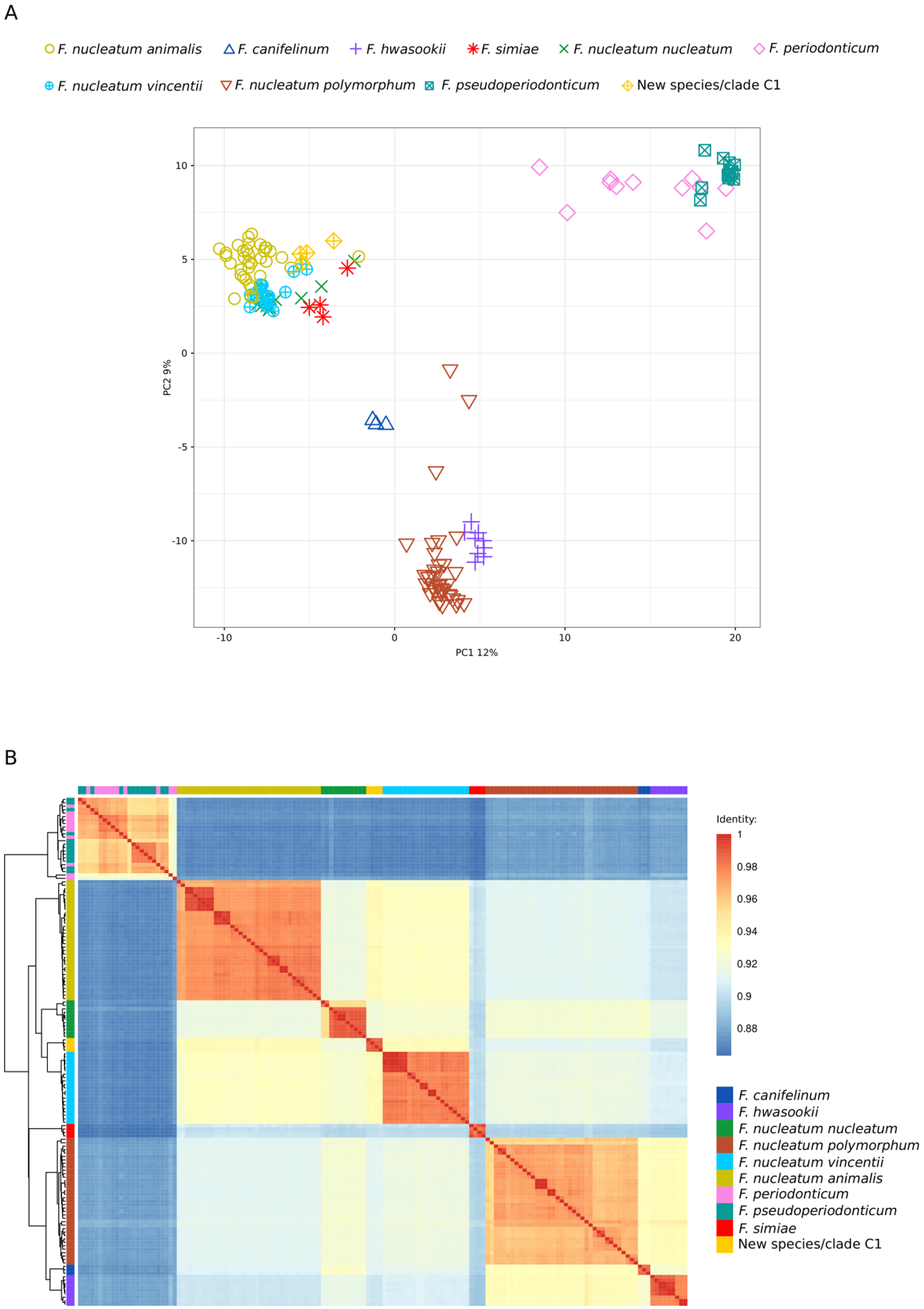

Fig. 7 | Analysis of accessory gene content and average nucleotide similarity (ANI). A PCA of accessory gene content. Each *Fusobacterium* genome is colored and displayed with a different symbol, as in Fig. 6. **B** ANI heatmap of full genomes. *Fusobacterium* species/subspecies are indicated with the same colors as in panel **A** and as in Fig. 6.

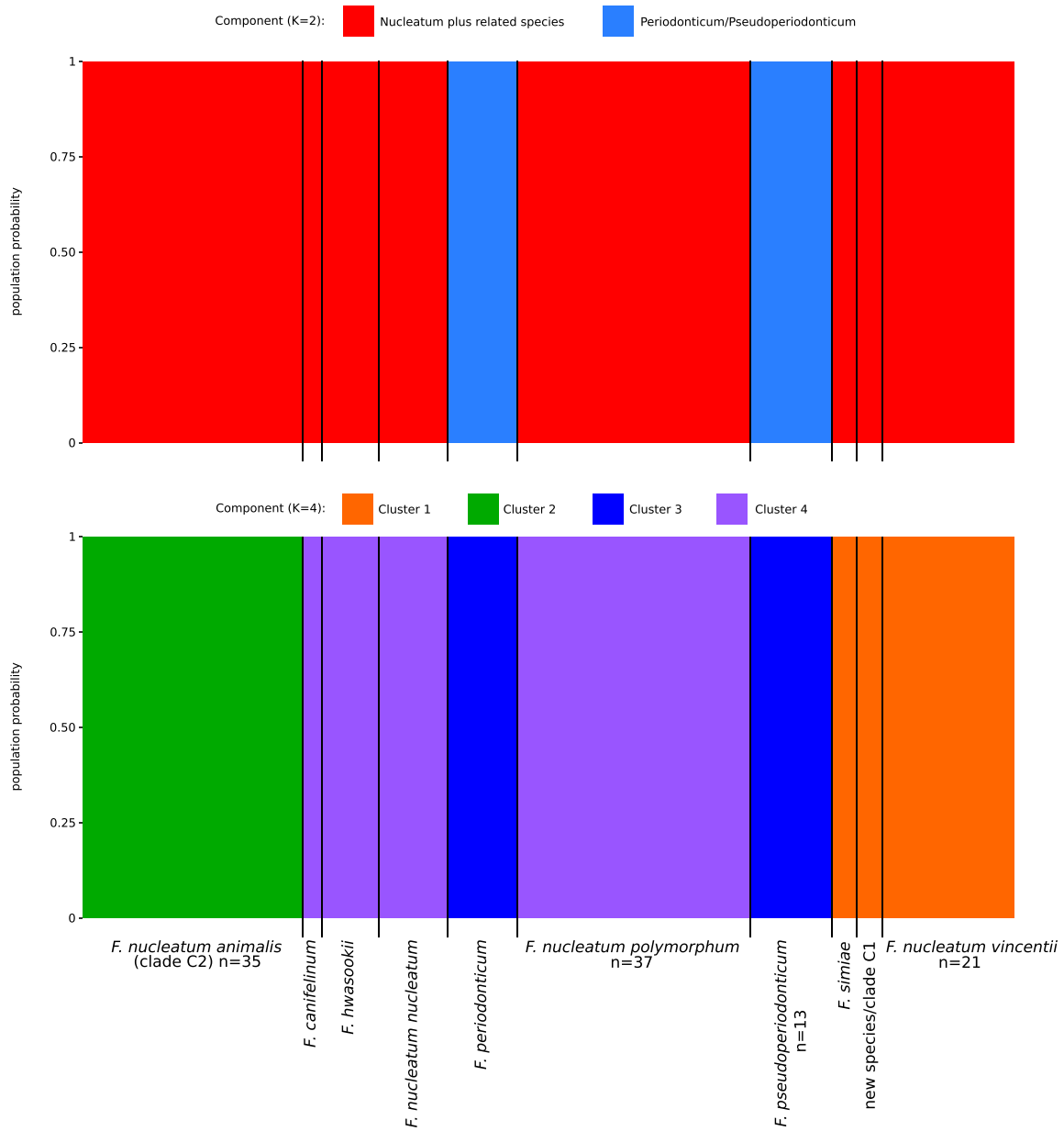

Fig. 8 | Population structure analysis: no admixture model. Bar plot representing the probability of population assignment from the STRUCTURE no admixture model. Each vertical line represents a *Fusobacterium* core genome. Results are shown for $K = 2$ and $K = 4$. For the latter, populations are colored as the cluster in PCA analysis (Fig. 6A).

association analyses. We thus reasoned that a comprehensive investigation of the genetic diversity and relationships within the *Fusobacterium* genus might provide valuable information for future investigation to contextualize disease associations.

We first asked if and how intensely fusobacteria recombine, and whether recombination can affect phylogenetic relationships. Indeed, one of the effects of recombination is to unlink loci along the genome, so that they evolve independently and display diverse evolutionary histories. We report both ancestral (occurring during speciation) and recent (occurring after speciation) recombination between species in *rpoB*, which is commonly used as a marker gene and was proposed by Bi and co-workers for sample profiling[12]. As a consequence, different gene regions show distinct patterns of sequence similarity among species. In general, we found evidence of recent and ancestral recombination in all core genes we analyzed, supporting the view that fusobacteria have mosaic genomes[25,26] and emphasizing that species/subspecies identification should not rely on single genes. In this respect, it is worth noting that a recent study showed that, in fusobacteria, specific gene families, such as adhesins, may undergo more extensive recombination and HGT than core genes[25]. Thus, our analyses may be biased in terms of overall recombination estimates, as only core genes were included. Nonetheless, core genes are the ones usually used for classification purposes, and this is the reason why we focused on them. The mechanism underlying the differences in the intensities of recombination among core genes will need further investigation. However, analyses in other bacterial species indicated that recombination rates are heterogeneous across the genome and they are influenced by local features such as distance from replication origin or proximity to mobile elements[38,39].

For a more detailed analysis, we focused on bacterial genomes in lineage 12 (from *rpoB* analysis), which show high sequence similarity in the analysis of core genes, as well. Although classified in different species and subspecies, these genomes are more closely related than most other fusobacteria and their relatively limited genetic diversity allows application of strategies to study population structure. Moreover, lineage 12 includes *Fusobacterium* species that have been intensively investigated

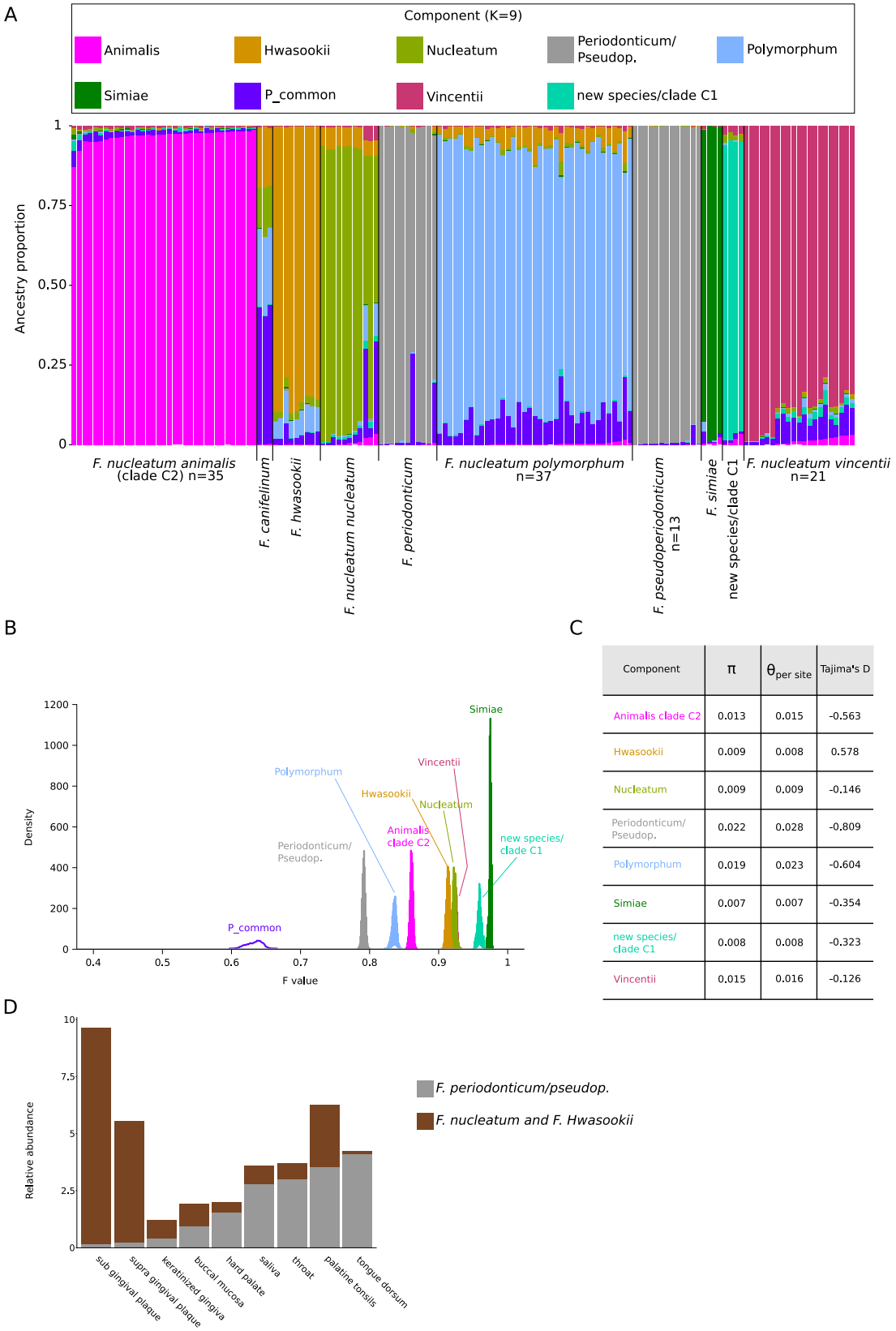

Fig. 9 | Population structure analysis: linkage model. A Bar plot representing the proportion of ancestral population components for $K = 9$. Each vertical line represents a *Fusobacterium* core genome and it is colored by the proportion of sites that have been assigned to the nine populations by STRUCTURE. Ancestry components are named based on the genomes where they are more prevalent. **B** Distributions of F values for the nine populations. Colors are as in panel **A**. **C** Nucleotide diversity and Tajima's for genomes that acquired a major part of their ancestry (> 80%) from individual populations. **D** Percent abundance of different *Fusobacterium* species in different oral sites (data were derived from the expanded Human Oral Microbiome Database). The four *F. nucleatum* subspecies were considered as a single group, to which *F. hwasookii* was also added.

for their role in CRC. Using the core genome, both the PCA and STRUCTURE analyses were consistent in showing that modern populations are divided into two sub-lineages, which comprise *F. periodonticum/pseudoperiodonticum* and *F. nucleatum* plus related species. Further grouping was however observed, with four major populations defined as clusters in the PCA and as modern sub-populations in the non admixture STRUCTURE model. In line with these data, ANI analysis of complete genomes identified at least nine species, with different levels of genetic similarity within and among themselves. Interestingly, an analysis of accessory genes confirmed the separation of *F. periodonticum/pseudoperiodonticum* from other species, but tended to cluster together *F. nucleatum nucleatum, F. nucleatum vincentii, F. nucleatum animalis*, and the new putative species/clade C1. Conversely, *F. nucleatum polymorphum* showed highest similarity in gene content to *F. hwasookii*. This is in agreement with a recent study that analyzed components of the type V secretion system[25]. Using an approach based on fastGEAR, Crowley and coworkers found that genes from *F. nucleatum nucleatum*, *animalis* and *vincentii*, tended to form a single lineage, which did not include genes from *F. nucleatum polymorphum*. As the authors suggested, their data and those reported herein might indicate more active horizontal gene transfer (HGT) among *F. nucleatum nucleatum*, *animalis*, and *vincentii* than with *F. nucleatum polymorphum*[25]. Whereas this might be due to the existence of different barriers to HGT (e.g., presence of similar/different restriction-modification systems), the separation of *F. periodonticum/pseudoperiodonticum* from the other species might also be caused by limited HGT due to the occupation of different oral niches.

As is the case of genomes in the wider collection of the *Fusobacterium* genus, miss-classification or incomplete taxomomic definition was common within lineage 12. However, using PCA or STRUCTURE analysis most genomes could be assigned to known species or subspecies. Nonetheless, *F. periodonticum* and *F. pseudoperiodonticum* could not be clearly differentiated using a range of methods (e.g., core genome PCA, phylogenetics, ANI, population structure analysis, accessory gene content). Thus, evidence herein does not substantiate the separation of these two species. Also, all analyses confirmed that four genomes that include strains FNU and 13-08-02 (in clade C1 in Zepeda-Rivera et al.) do not belong to the *F. nucleatum animalis* subspecies or any other known species/subspecies. Thus, our data do not support the previously suggested division of *F. nucleatum animalis* diversity in two clades[6]. In their recent study, Zepeda-Rivera and coworkers showed that clade C2 was associated to the CRC niche, whereas C1 was not. The two clades were suggested to differ in terms of accessory genome size, number of extrachromosomal plasmids and immune defences, as well as methylation patterns, representation of adhesins, and metabolic potential. The cells of bacteria in clade C2 were also found to be longer and thinner than those in clade C1, and to have a higher level of cancer cell invasion. All these differences would be noteworthy for bacteria in the same species, but we show here that this is not the case. Indeed, all these features were compared between two distinct species, although closely related. Whereas this is unlikely to change the conclusion that *F. nucleatum animalis* is associated with CRC, our data call for a re-assessment of the characteristics of clade C2, which were established in comparison to a different species[6]. This is particularly true in light of the PCA and STRUCTURE analyses, which indicated that the new specie/clade C1 and *F. nucleatum animalis* belong to different clusters or modern populations. Notably, the linkage model in STRUCTURE analysis showed that the new species/clade C1 inherited ancestry from a distinct ancestral population that experienced substantial drift in comparison to most other populations, including the one that contributed to the ancestry of *F. nucleatum animalis*. Genetic drift was shown to promote genome reduction and decreased coding density in bacteria[40], in line with the small genomes of clade C1 bacteria[6]. The new species/clade C1 also shows very limited genetic diversity, although analysis is not particularly robust, as it was necessarily limited to four genomes. Overall, these results indicate that more extensive and lineage-wise comparisons are necessary to establish which *F. nucleatum animalis* characteristics contribute to CRC association.

Interestingly, STRUCTURE results showed that the ancestral population from which *F. periodonticum/pseudoperiodonticum* emerged experienced the lowest drift and these species now display the highest genetic diversity. Together with the negative Tajima's D, this is suggestive of a size expansion in the population. Compared to other oral *Fusobacterium* species, *F. periodonticum* was found to commonly occur in different oral sites, with lower representation in the gingival plaque. Conversely, *F. nucleatum* subspecies and *F. hwasookii* were mostly specialized for plaque. These results are consistent with the view that, during their evolution *F. nucleatum* subpopulations drifted away from a common ancestral population to colonize a new niche.

Our work has limitations. One of the most serious concerns the scant meta-data available for the *Fusobacterium* genomes we analyzed. For most of them, we had no information about the origin in terms of geographic location, body site, or host. In most cases, the health status of the host was unknown, as well as the isolation source. Whereas this limits the possibility to perform a more detailed analysis of genetic diversity in fusobacteria, the inability to control for external variables might also introduce unrecognized biases. For instance, individual bacterial species were shown to be more genetically diverse among African than non-African human hosts[34,41–44]. The unequal representation of genomes from different geographic areas in different species might thus affect our measures of nucleotide diversity. Another limitation concerns our focus on core genomes, which was motivated by the need to obtain reliable alignments and PI sites, as well to maintain data tractable for STRUCTURE analysis. Although, bacterial pangenomes are known to be highly diverse and virulence factors are often encoded by accessory genes, the purpose of our work was to describe the genetic relationships in the *Fusobacterium* genus. We consider that these data might be valuable to develop a much needed molecular profiling approach that can shed light into the epidemiology of fusobacteria-associated diseases.

Methods

Bacterial and core genes sequences

The list of *Fusobacterium* genomes was derived from the BV-BRC site (https://www.bv-brc.org/, as of July 2023) by selecting entries with "good" genome quality. Complete and draft genome sequences were obtained from the NCBI database by using the getGenome function from the R package biomaRt[45]; the set consisted of 361 bacterial samples (Supplementary Data 1). Complete and metagenome-assembled genomes were used as input data for the Genome Taxonomy Database Toolkit (GTDB-Tk)[46]. This tool provides an automated taxonomic classification of bacterial sequences based on a set of 120 single copy marker proteins; GTDB-Tk also identifies and extracts from input genomes both nucleotide and protein sequences of each marker. The nucleotide sequences were then used for the subsequent analyses. Since not all samples have the whole genome covered, we were unable to retrieve all markers for all samples. For instance, the 16 genomes in which we failed to identify rpoB had either relatively low coverage or were assembled in a large number of contigs (Supplementary Data 1).

rpoB gene alignment and network

A nucleotide alignment based on 345 *rpoB* gene sequences was constructed using MAFFT with default parameters[47]. A neighbor-net split network was generated throughout SplitsTree4[48]: a data matrix was generated from the aligned sequences, estimating distances with the HKY85 model and removing parsimony-uninformative and constant sites.

Recombination and sequence similarity analyses

The same rpoB alignment described above was used to run fastGEAR, an algorithm that detects recombination events between inferred lineages, as well as from external origins. In particular, this method first clusters sequences into lineages, then it identifies both recent (i.e. affecting a subset of strains in a lineage), and ancestral (i.e. affecting all strains in a lineage) recombination events[49]. The same approach was used to identify ancestral and recent recombination events for a list of 45 genes from the 120 marker

genes. These 45 genes were selected because they were longer than 1000 nucleotides and they were present in at least 300 genomes (Supplementary Data 1). FastGEAR was run using the default settings and the output was then used to generate a plot of recent recombination events versus ancestral ones, normalized by gene alignment length.

A concatenated alignment, based on 120 core genes, was generated with the same genomes used in the rpoB alignment. The alignment was generated using the GUIDANCE2 suite[50], setting sequence type as nucleotides and using MAFFT as an aligner. GUIDANCE2 also allows to filter unreliably aligned positions. We thus removed positions with a score lower than 0.90[51].

Sequence similarity analyses were performed using SimPlot++[52]. This tool generates a similarity network plot based on a multisequence alignment. Each node of the network represents a sequence or group of sequences and edges indicate the global (over the whole sequence) or local (over one or more sub-regions) similarity among nodes.

Accessory gene identification and ANI analysis

Complete and draft genomes were annotated using Prokka[53]. Prokka uses Prodigal to identify potential genes/proteins present within the genome, then it compares these candidate genes with different databases, retrieving the annotation from the best match[54]. The general feature format (GFF) output of Prokka was used as input for PATO (Pangenome Analysis Toolkit) analysis[29], an R package that implements functions to run several external softwares, in order to perform advanced pangenome analyses. PATO was thus used to identify accessory genes within the lineage 12 strains; the Mmseqs function, which is a wrapper of the MMseq2 tool, was applied to search and cluster similar gene sequences. MMseq2 was used with clustering mode set to 0 (Greedy Set cover algorithm). After that, the mmseq object was used to run the accnet function, which builds a matrix containing the frequency of each accessory gene in the genome dataset: a gene was considered accessory if it had a maximum frequency of 0.8 in the lineage. Principal component analyses (PCA) was carried out using the matrix of gene presence/absence generated by PATO[29] and the mixOmics R package[55].

Whole genome average nucleotide identities were calculated for the strains belonging to lineage 12 using Pyani (v.0.2.12), a python module for whole genome classification of microbes[56]. In particular, the analysis was performed using ANIm[57], that is based on a MUMmer aligner[58]. Results were shown as a heatmap plot, using the pheatmap R package (https://cran.r-project.org/package=pheatmap), applying the "complete linkage" method as clustering algorithm.

PCA, population structure, and nucleotide diversity

Strains belonging to lineage 12 were selected to build a new concatenated gene alignment. Concatenated gene sequences that were shorter than 80% of the longest sequence were discarded from the analyses: this filtering allowed us to limit the number of gaps in the alignment but also to take into account differences in gene lengths; this generated a set of 148 strains. We then generated an alignment by applying GUIDANCE2 as described above. From this new alignment, biallelic (97% of the total) parsimony-informative (PI) sites were extracted; in particular, we selected biallelic sites, each with a minimum frequency of two, for those genomic positions where at least 50% of sequences had non-missing information. Gaps and all nonstandard nucleotide bases were considered as missing values. This generated a list of 26,430 variable positions. Principal component analysis (PCA) was performed with the mixOmics R package[55], using the PI matrix as input. The 3D PCA plot was generated with scatterplot3d R package[59]. The same PI data was also used to run STRUCTURE. First, the software was run with $K = 1$ to estimate the frequency spectrum parameter (λ), as suggested[32]. The λ parameter was estimated to be equal to 0.5878. Using this value, both the no admixture model with independent allele frequencies and the linkage model with correlated allele frequencies were run[31,32]. Both models were run with different values of K populations, from 1 to 12. To obtain more accurate inferences in spite of the different representation of genomes from distinct species/subspecies, we used an ancestry prior that allows source populations to contribute deferentially to the pooled sample of individuals[60]. In particular, for each *K*, 10 runs with a MCMC total chain length of 500,000 iterations and 50,000 iterations as burn-in were run. The optimal K was evaluated with Evanno's method[61] using the HARVESTER tool[62]. The CLUMPAK[63] software was used to combine replicate runs from the same K and to generate the Q value matrix. For the linkage model analysis, the amount of drift that each subpopulation experienced from a hypothetical ancestral population was quantified by the F parameter calculated for the optimal *k* value[32].

Finally, results obtained from the linkage model were used to group strains to estimate population genetic parameters. Specifically, each strain was assigned to one of the defined K populations if it had an ancestry component higher than 80% for that specific population (i.e. admixed individuals were excluded); then nucleotide diversity and Tajima's D were calculated for each populations using the DnaSP software[64].

Recombination-aware phylogenetic reconstruction

Phylogenetic trees were constructed by filtering recombinant regions using Gubbins v3.3.5 with default settings[28]. Gubbins generated a recombination-free alignment of polymorphic sites that was used as input for the fastTree[65] tool implemented in Gubbins with GTRGAMMA as the nucleotide substitution model. A co-phylogenetic plot between recombination-free trees of rpoB and concatenated alignments for the the 120 core genes was generated using the phytools R package[66]. Tree nodes rotation was allowed to optimize tip matching.

Relative abundance data

Bacterial percent abundance (average relative abundance of each oligotype in each district) in the human mouth and aerodigestive tract was retrieved from the expanded Human Oral Microbiome Database v3.1 (eHOMD) (https://www.homd.org/)[67]. In particular, we retrieved the percentage abundance of F. periodonticum, F. nucleatum (all subspecies), and F. hwasookii from three different experiments, as available in (eHOMD). We next summed the percentage abundance for F. nucleatum and F. hwasookii in each study. Finally, we calculated mean values for 9 different oral districts: buccal mucosa, keratinized gingiva, hard palate, tongue dorsum, palatine tonsils, throat, saliva, supra-gingival plaque, and sub-gingival plaque (Supplementary Table 2).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All Fusobacterium strains genbank accession IDs are listed in Supplementary Data 1. Source Data (alignments and phylogenetic trees) are available in the following Figshare repository: https://doi.org/10.6084/m9.figshare.26879839.v1.

Received: 22 April 2024; Accepted: 2 September 2024;
Published online: 07 September 2024

References

1. Brennan, C. A. & Garrett, W. S. Fusobacterium nucleatum - symbiont, opportunist and oncobacterium. *Nat. Rev. Microbiol.* **17**, 156–166 (2019).
2. Kostic, A. D. et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
3. Bullman, S. et al. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science* **358**, 1443–1448 (2017).
4. Barot, S. V. et al. Distinct intratumoral microbiome of young-onset and average-onset colorectal cancer. *EBioMedicine* **100**, 104980 (2024).

5. Younginger, B. S. et al. Enrichment of oral-derived bacteria in inflamed colorectal tumors and distinct associations of Fusobacterium in the mesenchymal subtype. *Cell Rep. Med.* **4**, 100920 (2023).
6. Zepeda-Rivera, M. et al. A distinct Fusobacterium nucleatum clade dominates the colorectal cancer niche. *Nature* **628**, 424–432 (2024).
7. Bučević Popović, V. et al. The urinary microbiome associated with bladder cancer. *Sci. Rep.* **8**, 12157 (2018).
8. Parhi, L. et al. Placental colonization by Fusobacterium nucleatum is mediated by binding of the Fap2 lectin to placentally displayed Gal-GalNAc. *Cell Rep.* **38**, 110537 (2022).
9. Vander Haar, E. L., So, J., Gyamfi-Bannerman, C. & Han, Y. W. Fusobacterium nucleatum and adverse pregnancy outcomes: Epidemiological and mechanistic evidence. *Anaerobe* **50**, 55–59 (2018).
10. Yeoh, Y. K. et al. Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut* **69**, 1998–2007 (2020).
11. He, Y. et al. Non-nucleatum Fusobacterium species are dominant in the Southern Chinese population with distinctive correlations to host diseases compared with F. nucleatum. *Gut* **70**, 810–812 (2021).
12. Bi, D. et al. Profiling Fusobacterium infection at high taxonomic resolution reveals lineage-specific correlations in colorectal cancer. *Nat. Commun.* **13**, 3336 (2022).
13. Tran, H. N. H. et al. Tumour microbiomes and Fusobacterium genomics in Vietnamese colorectal cancer patients. *NPJ Biofilms Microbiomes* **8**, 87 (2022).
14. Kook, J.-K. et al. Genome-based reclassification of fusobacterium nucleatum subspecies at the species level. *Curr. Microbiol.* **74**, 1137–1147 (2017).
15. Manson McGuire, A. et al. Evolution of invasion in a diverse set of Fusobacterium species. *mBio* **5**, e01864 (2014).
16. Ma, X. et al. Pangenomic study of fusobacterium nucleatum reveals the distribution of pathogenic genes and functional clusters at the subspecies and strain levels. *Microbiol. Spectr.* **11**, e0518422 (2023).
17. Preska Steinberg, A., Lin, M. & Kussell, E. Core genes can have higher recombination rates than accessory genes within global microbial populations. *Elife* **11**, e78533 (2022).
18. Sakoparnig, T., Field, C. & van Nimwegen, E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *Elife* **10**, e65366 (2021).
19. Shoemaker, W. R., Chen, D. & Garud, N. R. Comparative population genetics in the human gut microbiome. *Genome Biol. Evol.* **14**, evab116 (2022).
20. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
21. Lin, M. & Kussell, E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods* **16**, 199–204 (2019).
22. Crits-Christoph, A., Olm, M. R., Diamond, S., Bouma-Gregson, K. & Banfield, J. F. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J.* **14**, 1834–1846 (2020).
23. Stott, C. M. & Bobay, L.-M. Impact of homologous recombination on core genome phylogenies. *BMC Genomics* **21**, 829 (2020).
24. 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**, e0029722 (2022).
25. Crowley, C., Selvaraj, A., Hariharan, A., Healy, C. M. & Moran, G. P. Fusobacterium nucleatum subsp. polymorphum recovered from malignant and potentially malignant oral disease exhibit heterogeneity in adhesion phenotypes and adhesin gene copy number, shaped by inter-subspecies horizontal gene transfer and recombination-derived mosaicism. *Microb. Genomics* **10**, 001217 (2024).
26. Mira, A., Pushker, R., Legault, B. A., Moreira, D. & Rodríguez-Valera, F. Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics. *BMC Evol. Biol.* **4**, 50 (2004).
27. Fatahi-Bafghi, M. Genomic and phylogenomic analysis of Fusobacteriaceae family and proposal to reclassify Fusobacterium naviforme Jungano 1909 into a novel genus as Zandiella naviformis gen. nov., comb. nov. and reclassification of Fusobacterium necrophorum subsp. funduliforme as later heterotypic synonym of Fusobacterium necrophorum subsp. necrophorum and Fusobacterium equinum as later heterotypic synonym of Fusobacterium gonidiaformans. *Antonie van. Leeuwenhoek* **117**, 34 (2024).
28. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
29. Fernández-de-Bobadilla, M. D. et al. PATO: pangenome analysis toolkit. *Bioinformatics* **37**, 4564–4566 (2021).
30. Ciufo, S. et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* **68**, 2386–2392 (2018).
31. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
32. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
33. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).
34. Falush, D. et al. Traces of human migrations in Helicobacter pylori populations. *Science* **299**, 1582–1585 (2003).
35. Escapa, I. F. et al. New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems* **3**, e00187–18 (2018).
36. Eren, A. M., Borisy, G. G., Huse, S. M. & Mark Welch, J. L. Oligotyping analysis of the human oral microbiome. *Proc. Natl Acad. Sci. USA* **111**, E2875–E2884 (2014).
37. Ye, X. et al. *Fusobacterium Nucleatum* subspecies *Animalis* influences proinflammatory cytokine expression and monocyte activation in human colorectal tumors. *Cancer Prev. Res.* **10**, 398–409 (2017).
38. Didelot, X., Méric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics* **13**, 256 (2012).
39. Everitt, R. G. et al. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nat. Commun.* **5**, 3956 (2014).
40. Kuo, C.-H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
41. Mah, J. C., Lohmueller, K. E. & Garud, N. Inference of the demographic histories and selective effects of human gut commensal microbiota over the course of human history. *bioRxiv* https://doi.org/10.1101/2023.11.09.566454 (2023).
42. Tett, A. et al. The Prevotella copri complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe* **26**, 666–679.e7 (2019).
43. Karcher, N. et al. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biol.* **21**, 138 (2020).
44. Schnorr, S. L. et al. Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654 (2014).

45. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
46. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
48. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
49. Mostowy, R. et al. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol. Biol. Evol.* **34**, 1167–1182 (2017).
50. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* **43**, 7 (2015).
51. Privman, E., Penn, O. & Pupko, T. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.* **29**, 1–5 (2012).
52. Samson, S., Lord, É. & Makarenkov, V. SimPlot++: a Python application for representing sequence similarity and detecting recombination. *Bioinformatics* **38**, 3118–3120 (2022).
53. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
54. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
55. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
56. 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**, 12–24 (2016).
57. Adler, A., Poirier, S., Pagni, M., Maillard, J. & Holliger, C. Disentangle genus microdiversity within a complex microbial community by using a multi-distance long-read binning method: example of *Candidatus* Accumulibacter. *Environ. Microbiol.* **24**, 2136–2156 (2022).
58. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
59. Ligges, U. & Mächler, M. **scatterplot3d** - An R Package for Visualizing Multivariate Data. *J. Stat. Soft*. **8**, 1–20 (2003).
60. Wang, J. The computer program structure for assigning individuals to populations: easy to use but easier to misuse. *Mol. Ecol. Resour.* **17**, 981–990 (2017).
61. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
62. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet Resour.* **4**, 359–361 (2012).
63. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15**, 1179–1191 (2015).
64. Rozas, J. et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).
65. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
66. Revell, L. J. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* **12**, e16505 (2024).
67. Chen, T. et al. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010**, baq013 (2010).

Acknowledgements

This work was supported by the Italian Ministry of Health ("Ricerca Corrente" to M.S.).

Author contributions

Conceptualization, M.S. and D.F.; Methodology, M.S., C.M., and D.F.; Investigation, C.M., D.F., and R.C.; Writing Original Draft, M.S. and C.M.; Writing review & editing, M.S., D.F., and R.C.; Funding Acquisition, M.S.

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/s42003-024-06825-y.

Correspondence and requests for materials should be addressed to Cristian Molteni.

Peer review information *Communications Biology* thanks Gary Moran and Hao Chung Theand for their contribution to the peer review of this work. Primary Handling Editors: Pei Hao and Tobias Goris.