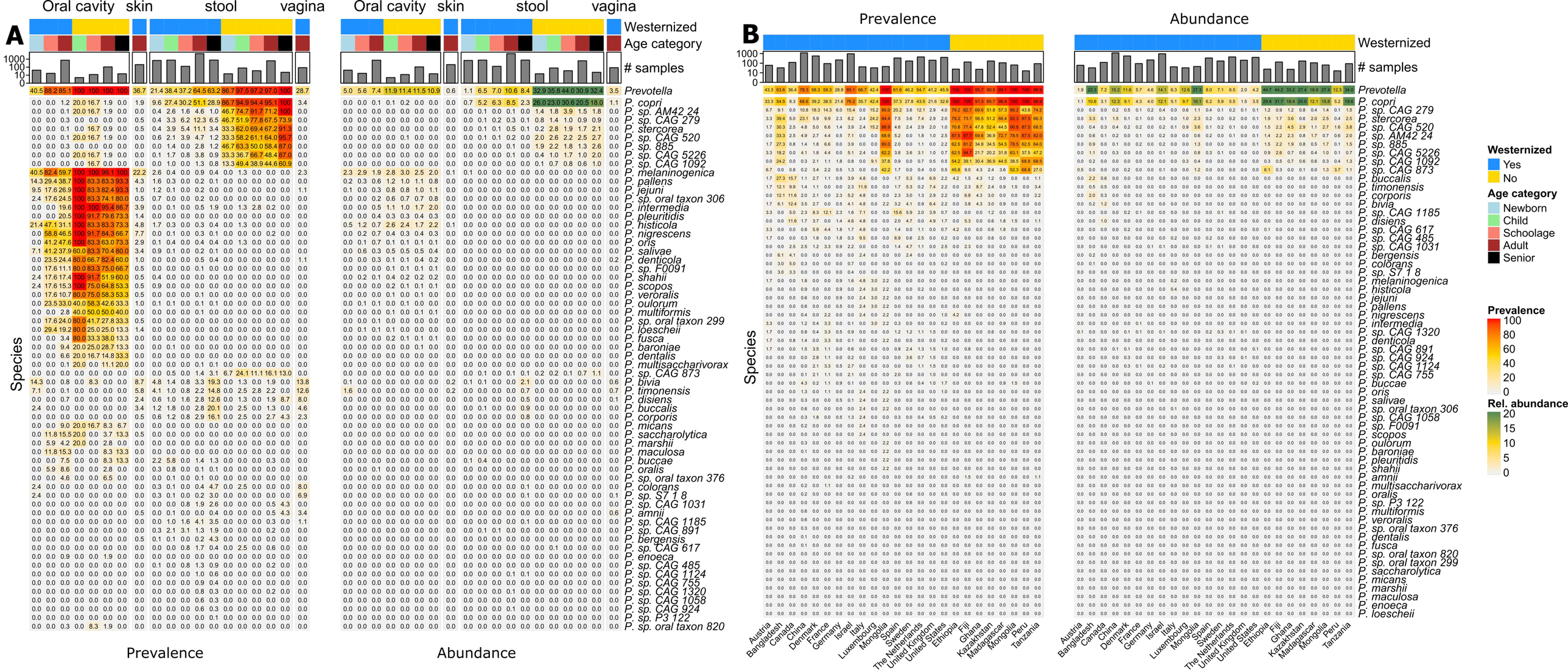
nature portfolio

https://doi.org/10.1038/s41579-021-00559-y

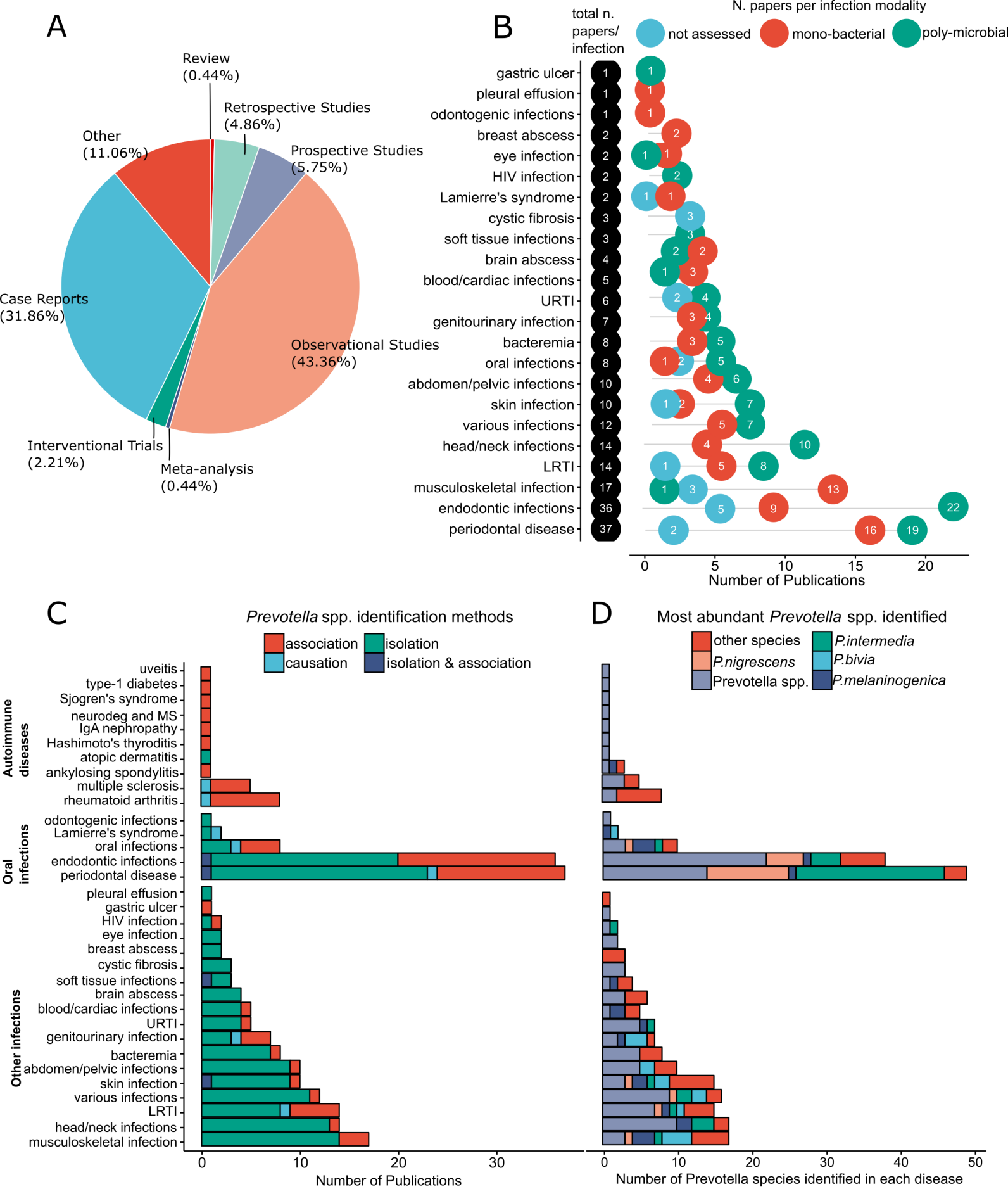
Supplementary information

Prevotella diversity, niches and interactions with the human host

In the format provided by the authors and unedited



Supplementary Figure 1. Distribution and stratification of *Prevotella* species across human populations and body-sites. Prevalence and relative abundance of *Prevotella* species across **(A)** body-sites, age and lifestyles, and **(B)** countries by exploiting public integrated repositories of >9,500 profiled human metagenomes. The heatmaps extend the results reported in **Figure 2** by including also unnamed species.



Supplementary Figure 2. An overview of the identified links between *Prevotella* species and human diseases. **(A)** As a result of the systematic literature review (see Supplementary Information box 4) on the role of *Prevotella* spp. in human infections, we surveyed 226 papers that are here classified according to the type of the study proposed (%). **(B)** Total number of papers per infection or disease surveyed, and number of papers in each disease related to mono-bacterial or poly-microbial infection types. URTI, upper-respiratory tract infections. LRTI, low-respiratory tract infections. NA, not clearly assessed in the study. **(C)** Number of publications per disease according to the methodology used to identify the presence of *Prevotella* spp. Isolation refers to standard microbiology approaches of isolation and cultivation. Association refers to low- or high-input molecular approaches such as species-specific PCR or through 16S rRNA gene sequencing. Causation refers only to studies where the contribution of *Prevotella* spp. was proved in animal models. **(D)** Number of the top-4 most abundant *Prevotella* species identified in each disease investigated. Low-abundant species were grouped into "other species", while *Prevotella* spp. was used when the species was not identified or reported.

Supplementary box 1. *Prevotella* isolate diversity.

For **Figure 1**, **Table 1** and **Supplementary Data 1**, we considered the genomes labelled as "Prevotellaceae" available in the NCBI Reference Sequence Database (RefSeq) as of May 2020 (*n* = 384, **Table 1**). Metadata information was retrieved directly from the NCBI portal or, when missing, from related publications. Genomes were dereplicated at 0.5% genetic distance based on whole-genome nucleotide similarity estimating using Mash (v. 2.0; option "-s 10000" for sketching)[1]. Dereplicated genomes (*n* = 254) were considered to build the phylogenetic tree of strains of *Prevotella* and other *Prevotellaceae* species using PhyloPhlAn 3.0 (ref.[2]) and GraPhlAn[3] (**Figure 1a**). Genomes were annotated with Prokka (v. 1.12)[4] using default parameters. Proteins inferred by Prokka were then functionally annotated with UniRef90 using diamond (v. 0.9.9.110)[5]. Statistics on the coding regions (CDS) were reported for the species with at least five genomes (**Figure 1b**).

Supplementary box 2. Description of metagenomic datasets.

For **Figure 2**, we considered 9,539 human metagenomes coming from 51 publicly available independent datasets and with curated metadata information available in ref.[6]. Taxonomic profiles were generated using MetaPhlAn 3.0[7]. Prevalence and relative abundance across multiple host conditions (i.e., age category, body-site, country, and lifestyle) are reported only for named *Prevotella* species with prevalence > 0.1% in at least one category in **Figure 2a-c**. Extended results by including also "*Prevotella sp."* species are reported in **Supplementary Figure 1**.

Supplementary box 3. Prevotella MAG diversity.

For **Figure 3** and **Supplementary Data 1**, we considered 213,635 MAGs that were retrieved from publicly available metagenomes coming from human[8], non-human primate[9], food[10], and other hosts that we retrieved from refs[11,12] or directly reconstructed from raw reads deposited in NCBI SRA. We kept only MAGs having completeness > 50% and contamination < 5% as estimated by CheckM[13]. Such MAGs were integrated with the set of 384 reference genomes (summarized in **Table 1**) and clustered into SGBs following the procedure proposed in ref.[8]. Genomes were clustered with average linkage at 5% genetic distance based on whole-genome nucleotide similarity estimation using Mash (v. 2.0; option "-s 10000" for sketching)[1]. We identified a total of 10,243 genomes (384 reference genomes and 9,859 MAGs; **Supplementary Data 1**) belonging to *Prevotellaceae* and coming from different sources: dog (*n* = 2), elephant (*n* = 1), environmental (*n* = 4), hoatzin (*n* = 1) human (*n* = 7,819), mouse (*n* = 142), non-human primate (*n* = 214), ruminant (*n* = 2,041), and swine (*n* = 19). Such genomes were grouped into 526 distinct SGBs. We considered the 56 most prevalent human *Prevotella* SGBs (i.e., with at least twenty genomes retrieved from human

microbiomes) to build the phylogenetic tree of **Figure 3a**, in which only fifteen randomly selected genomes were considered for each SGB. The same set of SGBs was taken into account in **Figure 3b**.

Supplementary box 4. Prevotella as a potential pathogen.

For **Figure 4**, **Supplementary Figure 2** and **Supplementary Data 2**: The systematic review of the literature was conducted searching for "*Prevotella* & Infection & Humans" in Pubmed (June 2020). A total of 2,225 results were initially obtained. Exclusion criteria for study selection were: i) absence of the *Prevotella* spp., ii) healthy conditions, iii) non-infectious specimens, iv) broad reviews citing other main findings. After applying those criteria, 205 infections-related papers were retained and investigated. Another 21 autoimmune diseases, already discussed in other paragraphs, were also surveyed in a similar manner. Infections were initially grouped into broad categories such as "other infections" (accounting for the 54% of the total) or "oral infections" (37%), while the remaining 9% is the autoimmune disease category. **Supplementary Data 2** includes all the data extracted from the 226 papers analysed and used in this work. Statistical analysis and figures were generated in R from the information stored in **Supplementary Data 2**. **Fig. 4a** represents the associations between each *Prevotella* species surveyed and the corresponding disease (represented as broad classifications). The network was created with Cytoscape v3.8.1. and edge thickness represents the number of papers showing such an association between species and disease. For **Fig. 4b**, the methodology of *Prevotella* spp. identification were surveyed and displayed for each broad-classification category and in the most prevalent diseases observed (i.e. those described in at least 4 studies). "Isolation" refers to standard microbiology approaches. "Association" refers to low- or high-input molecular approaches such as species-specific PCR or through 16S rRNA gene sequencing. "Causation" refers only to studies where the contribution of *Prevotella* spp. was proved in animal models. The number of papers in each category were normalized against the total number of papers surveyed and showed as percentages. In **Supplementary Figure 2A**, papers were grouped by the type of study they proposed: the majority were observational studies (44%), followed by case-reports (32%), prospective (6%) and retrospective studies (5%). **Supplementary Figure 2B** shows the total number of papers per disease investigated (black circles) and the number of papers related to a "monobacterial" or "polymicrobial" nature of the *Prevotella* spp. infections. Studies that did not clearly state the type of infections were referred to as "not assessed". **Supplementary Figure 2C** includes the methodology used to identify the *Prevotella* species in each disease, as an extended data for Figure 4B**.** The top-4 most abundant *Prevotella* spp. identified in each disease were included in **Supplementary Figure 2D**, which included *P. intermedia*, *P. bivia*, *P. nigrescens* and *P. melaninogenica.* Low-

abundant species were grouped into “other species”, while “*Prevotella* spp.” was used when the species was not identified or reported.

References

1. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
2. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
3. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* vol. 3 e1029 (2015).
4. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
5. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
6. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
7. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
8. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
9. Manara, S. *et al.* Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biol.* **20**, 299 (2019).
10. Pasolli, E. *et al.* Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. *Nat. Commun.* **11**, 2610 (2020).
11. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
12. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).
13. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).