Evolutionary and Pan-genome Analysis of Three Important Black-pigmented Periodontal Pathogens

Pei Qi MENG[1], Qian ZHANG[2], Yun DING[1], Jiu Xiang LIN[3], Feng CHEN[2]

Objective: *To analyse the pan-genome of three black-pigmented periodontal pathogens: Porphyromonas gingivalis, Prevotella intermedia and Prevotella nigrescens.*
Methods: *Pan-genome analyses of 66, 33 and 5 publicly available whole-genome sequences of P. gingivalis, P. intermedia and P. nigrescens, respectively, were performed using Pan-genome Analysis Pipeline software (version 1.2.1; Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, PR China). Phylogenetic trees were constructed based on the entire pan-genome and single nucleotide polymorphisms within the core genome. The distribution and abundance of virulence genes in the core and dispensable genomes were also compared in the three species.*
Results: *All three species possess an open pan-genome. The core genome of P. gingivalis, P. intermedia and P. nigrescens included 1001, 1514 and 1745 orthologous groups, respectively, which were mainly related to basic cellular functions such as metabolism. The dispensable genome of P. gingivalis, P. intermedia and P. nigrescens was composed of 2814, 2689 and 906 orthologous groups, respectively, and it was enriched in genes involved in pathogenicity or with unknown functions. Phylogenetic trees presented a clear separation of P. gingivalis, P. intermedia and P. nigrescens, verifying the reclassification of the black-pigmented species. Furthermore, the three species shared almost the same virulence factors involved in adhesion, proteolysis and evasion of host defences. Some of these virulence genes were conserved across species whereas others belonged to the dispensable genome, which might be acquired through horizontal gene transfer.*
Conclusion: *This study highlighted the usefulness of pan-genome analysis to infer evolutionary cues for black-pigmented species, indicating their homology and phylogenomic diversity.*
Key words: *core genome, pan-genome, Porphyromonas gingivalis, Prevotella intermedia, Prevotella nigrescens*
Chin J Dent Res 2023;26(2):93–104; doi: 10.3290/j.cjdr.b4128023

Black-pigmented, gram-negative anaerobes have been found to be associated with periodontal disease since they were first isolated in 1928[1]. Given that these anaerobes share similar biochemical features and form black colonies on blood agar, they were named as *Bacteroides melaninogenica* prior to the 1990s. With advances in DNA sequencing, the underlying diversity within some of these black-pigmented anaerobes previously characterised as a single bacterial species became apparent. Through a systematic analysis using biochemical, physiological and immunological approaches, in addition to molecular DNA technologies, what was previously

1 Third Clinical Division, Peking University School and Hospital of Stomatology & National Clinical Research Center for Oral Diseases & National Engineering Laboratory for Digital and Material Technology of Stomatology & Beijing Key Laboratory of Digital Stomatology, Beijing, P.R. China.
2 Central Laboratory, Peking University School and Hospital of Stomatology & National Clinical Research Center for Oral Diseases & National Engineering Laboratory for Digital and Material Technology of Stomatology & Beijing Key Laboratory of Digital Stomatology, Beijing, P.R. China.
3 Department of Orthodontics, Peking University School and Hospital of Stomatology & National Clinical Research Center for Oral Diseases & National Engineering Laboratory for Digital and Material Technology of Stomatology & Beijing Key Laboratory of Digital Stomatology, Beijing, P.R. China.

Corresponding author: Dr Feng CHEN, Central Laboratory, Peking University School and Hospital of Stomatology, 22 Zhongguancun South Avenue, Haidian District, Beijing 100081, P.R. China. Tel: 86-10-82195773; Fax: 86-10-82195773. Email: chenfeng2011@hsc.pku.edu.cn

This work was supported by the National Natural Science Foundation of China (Grant No. 81991501), the KCL and PKUHSC Joint Institute for Medical Research Fund (Grant No. BMU2020KCL003) and the Post-doctor Seed Funding from Peking University School and Hospital of Stomatology.

a single entity is now appreciated to contain two different genera, *porphyromonas* and *prevotella*[2]. Of the many species of *porphyromonas* and *prevotella*, *Porphyromonas gingivalis* (*P. gingivalis*) and *Prevotella intermedia* (*P. intermedia*) are considered major pathogens in periodontal infections in humans. Meanwhile, the heterogeneity within *P. intermedia* species has led to further biochemical and chemical studies, which have revealed two serotypes within the *P. intermedia* species. The newly discovered *P. intermedia*, which had another species identified within this assumed homogenous species, is what we now know as *Prevotella nigrescens* (*P. nigrescens*)[3]. *P. nigrescens*, which is also considered a periodontal pathogen, causes periodontitis by colonising on the tooth surface and triggering an overly aggressive immune response, resulting in inflammation of oral tissue.

In the last few decades, the development of high-throughput sequencing technology and bioinformatics has increased the availability of sequenced bacterial genomes. The first published whole-genome sequence of *P. gingivalis* and *P. intermedia* was *P. gingivalis* W83[4] and *P. intermedia* 17[5], respectively. Since then, several whole-genome sequences covering various species of *porphyromonas* and *prevotella* genera have become available in the public domain. Analyses of the available genome data have the potential to provide insights into the evolution and phylogeny of these closely correlated black-pigmented strains.

The accumulation of genome sequences introduced the concept of a "pan-genome"[6]. The pan-genome is defined as the set of all the genes presented in an analysed dataset and comprises both core and dispensable genomes[7]. The core genome refers to the set of genes shared by almost all the genomes of the analysed dataset, whereas the dispensable genome describes the genes shared within only one or some genomes. Since the original proposal of the pan-genome, this concept has been widely used to account for the genomic diversity present within a given phylogenetic clade, including the heterogeneous genus *lactobacillus*[8,9], *clostridium*[10], and the species *escherichia coli*[11], *pseudomonas aeruginosa*[12] and *streptococcus agalactiae*[6].

However, despite the availability of several whole-genome sequences of black-pigmented species in public databases, no studies on the pan-genomics of these species are available. Although a few recent studies investigated the genomics of *P. gingivalis* and *P. intermedia*[13-16], none performed comprehensive pan-genome analyses on these species. Thus, in the present study, the authors performed a large-scale, comparative genome and pan-genome analysis of black-pigmented species (*P. gingivalis*, *P. intermedia* and *P. nigrescens*) using the publicly available whole-genome sequences of these species. Such genome-based analysis provides a detailed overview of the gene content of the core genome and pan-genome, and offers insight into phylogenomic relationships of these closely related clinically important species.

Materials and methods

Genome sequences

All the publicly available genome sequences of *P. gingivalis*, *P. intermedia* and *P. nigrescens* were obtained from the genome database of National Center for Biotechnology Information (NCBI). A total of 104 genome (draft and complete) sequences, including 66 *P. gingivalis*, 33 *P. intermedia* and 5 *P. nigrescens* genome sequences, which were submitted to the database prior to 12 October 2020, were used in this study. In addition, 19 genome sequences from bacteria in the bacteroides genus that were isolated from the human oral cavity were also downloaded and used as an outgroup for comparison in our analyses. All the sequences were from humans originating from diverse countries (USA, UK, South Korea, China, Japan, Sudan, Germany, Norway, Chile, Canada, Romania, Russia, Denmark and Hungary).

Clustering and functional annotation of the core and dispensable genomes

The annotated protein sequences of 104 genomes were grouped into homogenous clusters using OrthoMCL (University of Pennsylvania, Philadelphia, PA, USA)[17] based on sequence similarity. The BLAST reciprocal best hit algorithm[18] was applied with the criterion of e-value < $1e^{-5}$, identity > 40% and length coverage of a gene > 50%, and Markov cluster algorithms19 were employed with an inflation index of 1.5 to complete cluster analyses.

The functional category of each homogenous cluster was determined by performing BLASTp[20] against the Cluster of Orthologous Groups (COG) database (http://www.ncbi.nih.gov/COG/) with a criterion of e-value < $1e^{-5}$ and identity > 40%. To elucidate whether the core genome was enriched in a particular function, the proportions of the COG categories in the core and dispensable genomes were compared. An enrichment analysis was performed with a chi-square test using SPSS version 20 (IBM, Armonk, NY, USA) to give statistical significance to the difference. $P < 0.05$ was considered significant.

Statistical estimation of core genome and pan-genome size

Analyses of the pan-genome and core genome were undertaken using PGAP (version 1.2.1) software platform[21] for *P. gingivalis*, *P. intermedia* and *P. nigrescens*. The pan-genome and core genome were calculated as described previously[22] in an additive and reductive manner. Considering that core genes may be missed during genome sequencing and assembly, a correction step was introduced for the calculation of core genome size, in which any one gene that was absent in only one of the draft genomes was still regarded as a core gene[23]. The number of total genes/core genes provided by each added new genome depends on the selection of previously added genomes. For a given number of given strains (N), the number of all possible combinations is C (N_{total}, N). N_{total} represents the total number of genomes for a species. In this study, N_{total} is 66, 33, and 5 for *P. gingivalis*, *P. intermedia* and *P. nigrescens*, respectively. If the value of C (N_{total}, N) was greater than 8000, only 8000 random combinations were used. If fewer, all possible combinations were used. The final size of the pan- or core genomes was the mean value of all used combinations.

To perform statistical extrapolation to estimate the theoretical pan-genome and core genome size, a non-linear least-squares curve was used to fit the observed core genome and pan-genome sizes as a function of the number of analysed genomes. For the core genome extrapolation, an exponential decay function was used[6] where n is the genome number, κ and τ are fitting parameters and Ω is the extrapolated size of the core genome when $n \to \infty$.

$$f(n) = \kappa exp(-n/\tau) + \Omega$$

For the pan-genome, a Heaps' power law function was used, where n is the number of genomes used, a and b are fitting parameters and c is the growth exponent that indicates the speed at which the pan-genome is growing[24].

$$f(n) = a + bn^c$$

Results were compared among P. *gingivalis*, *P. intermedia* and *P.* nigrescens. For visual comparison, genome development trend maps were generated using OriginLab software (https://www.originlab.com/) (Northampton, MA, USA).

Phylogenetic analysis

For phylogenetic analyses, the PGAP pipeline used both the pan-genome profile and single nucleotide polymorphisms (SNPs) information in the core genome. Every set of orthologous genes found in all of the genomes were aligned separately using the multiple alignment tool MUSCLE (University of California, Berkeley, CA, USA)[25]. SNPs were extracted from these alignments and concatenated to form a multiple sequence alignment. Based on the whole pan-genome and SNPs within the core genome, phylogenetic trees were constructed using various methods, including an unweighted pair-group method with arithmetic means (UPGMA), neighbour-joining (NJ) and maximum likelihood (ML).

Additionally, a phylogenetic tree was constructed based on the 16S ribosome RNA (16S rRNA) gene, which is the most widely used gene marker in bacterial phylogenetic analysis for its functionally conserved feature. The 16S rRNA sequences with a length between 1,400 and 1,700 nt and an RNAmmer score above 1,700 were identified using RNAmmer (University of Oslo, Oslo, Norway)[26]. MEGA-7.0 software (King Abdulaziz University, Jeddah, Saudi Arabia)[27] was used to align the 16S rRNA sequences and construct an NJ tree with the Kimura 2-parameter model. Bootstrap values of each branch were calculated 500 times. All the trees were visualised using Evolview[28] (Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, PR China) (http://www.evolgenius.info/evolview.html).

To further reaffirm that the genomes of each strain did indeed belong to the assigned species, average nucleotide identity (ANI) values were calculated for all strains using the popular JSpecies package (Institut Mediterrani d'Estudis Avançats, Esporles, Spain)[29]. The ANI-based all-vs-all matrix was presented in a heatmap using HemI software (Huazhong University of Science and Technology, Wuhan, PR China)[30] (http://hemi.biocuckoo.org/down.php).

Analysis of virulence factors

To explore the distribution of virulence factors in the black-pigmented bacterial species, we reviewed the available literature and summarised the virulence factors of *P. gingivalis*, *P. intermedia* and *P. nigrescens* from previous publications[15,31-36]. The genome annotation reports for each strain were downloaded from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/all/), and the authors searched for the relevant genes associated with these virulence factors. The distribution and abundance of virulence genes in 66 *P. gingivalis*, 33 *P. intermedia* and

5 *P. nigrescens* genomes were displayed as a heatmap using HemI software. The abundance of virulence genes in the core and dispensable genomes was also compared in *P. gingivalis*, *P. intermedia* and *P. nigrescens*.

Results and discussion

Distribution of homogenous clusters

Detailed information regarding the 66 *P. gingivalis*, 33 *P. intermedia* and 5 *P. nigrescens* strains is shown in Supplementary Table S1 (provided on request). The genome size of *P. nigrescens* (mean 2.85 Mb and median 2.84 Mb) and *P. intermedia* (mean 2.79 Mb and median 2.78 Mb) is larger than that of *P. gingivalis* (mean 2.32 Mb and median 2.33 Mb). The genomes also vary substantially with regard to their number of genes. The number of genes was lowest for *P. gingivalis* with a mean of 2073, followed by 2324 for *P. intermedia* and 2407 for *P. nigrescens*; however, all three species have a relatively consistent genomic guanine and cytosine (GC) content, with *P. gingivalis* having the highest (range 48.1% to 49.1%), followed by *P. intermedia* (range 43.2% to 43.9%) and *P. nigrescens* (range 42.5% to 42.8%). By comparing the proteins annotated in the genomes of three species, 3815, 4203 and 2651 homologous clusters were identified in *P. gingivalis*, *P. intermedia* and *P. nigrescens*, respectively. Among these, 1001, 1514 and 1745 clusters comprised the core genome, while the remaining 2814, 2689 and 906 clusters comprised the dispensable genome (Table 1).

To identify functional differences between genes in the core and dispensable genomes, genes were classified according to their predicted function based on COG categories. The abundance of each COG category was plotted and compared in Figs 1a, 1c and 1e. In all three species, the highest numbers of genes in the core genome were related to translation, ribosomal structure and biogenesis (J). This was consistent with some species, such as *streptococcus pneumoniae*[37] and the genus *clostridium*[10], while other species, such as *staphylococcus aureus*, *salmonella enterica*, *escherichia coli*, *pseudomonas aeruginosa* and *acinetobacter baumannii*, showed amino acid transport and metabolism (E) as the most abundant category in their core genome[37].

With the exception of the categories P (inorganic ion transport and metabolism) and Q (secondary metabolites biosynthesis, transport and catabolism) in *P. nigrescens*, the enriched genes in the core genome were involved in almost all the COG categories of metabolism, including energy production and conversion (C), carbohydrate transport and metabolism (G), amino acid transport and metabolism (E), nucleotide transport and metabolism (F), coenzyme transport and metabolism (H), and lipid transport and metabolism (I); however, the enriched genes in the dispensable genome were included in poorly characterised categories, such as genes with unknown functions (S) in *P. gingivalis*, and only general function prediction (R) in *P. intermedia* and *P. nigrescens*. Similar enrichment in genes with no assigned function was reported previously in the dispensable genome of other organisms[6,38,39], and these genes may represent new virulent factors or interaction features of the oral microbiome. For all three species, statistically significant differences between the core and dispensable genome were also found in COG categories J (translation, ribosomal structure and biogenesis), M (cell wall/membrane/envelope biogenesis), O (posttranslational modification, protein turnover, chaperones), V (defence mechanisms) and X (mobilome: prophages, transposons). The core genome was enriched in the former three categories, whereas the dispensable genome was enriched in the latter two. In addition, we only observed in *P. nigrescens* that genes involved in cell cycle control, cell division and chromosome partitioning (D) comprised a significantly higher proportion of the core genome compared to the dispensable genome. Thus, genes involved in basic functions in life maintenance of the organism were likely to be more abundant in the core genome, whereas those related to pathogenicity or unknown functions were more likely to be enriched in the dispensable genome.

To compare the evolutionary pressure of genes in the core genome and dispensable genome, we calculated the ratio of non-synonymous to synonymous substitutions (dn/ds) for *P. gingivalis*, *P. intermedia* and *P. nigrescens*. The dn/ds ratio is commonly regarded as one of the most popular and reliable measures of the strength and mode of natural selection acting on protein coding sequences, with dn/ds > 1 indicating positive (adaptive or diversifying) selection, dn/ds = 1 indicating neutral evolution and dn/ds < 1 indicating negative (purifying) selection[40]. In the present study, the mean value of dn/ds was 1.22, 1.43 and 1.24 for the dispensable genome of *P. gingivalis*, *P. intermedia* and *P. nigrescens*, respectively, which was significantly higher than that of the core genome ($P < 0.001$), whose dn/ds value was below 1 in all three species (Figs 1b, 1d and 1f). It can be inferred that core genes are more conserved since they are involved in basic functions of life maintenance, whereas in the dispensable genome, the non-synonymous rate is significantly higher than that of the synonymous rate, leading to adaptive protein evolution to survive sudden environmental changes.

Table 1 Summary of the pan-genome of *P. gingivalis*, *P. intermedia* and *P. nigrescens*.

Number of clusters of orthologous genes	*P. gingivalis*	*P. intermedia*	*P. nigrescens*
Mean number of genes	3,815	4,203	2,651
Core	1,001 (26%)	1,514 (36%)	1,745 (66%)
Dispensable	2,814 (74%)	2,689 (64%)	906 (34%)
Strain-specific	759 (20%)	815 (19%)	488 (18%)

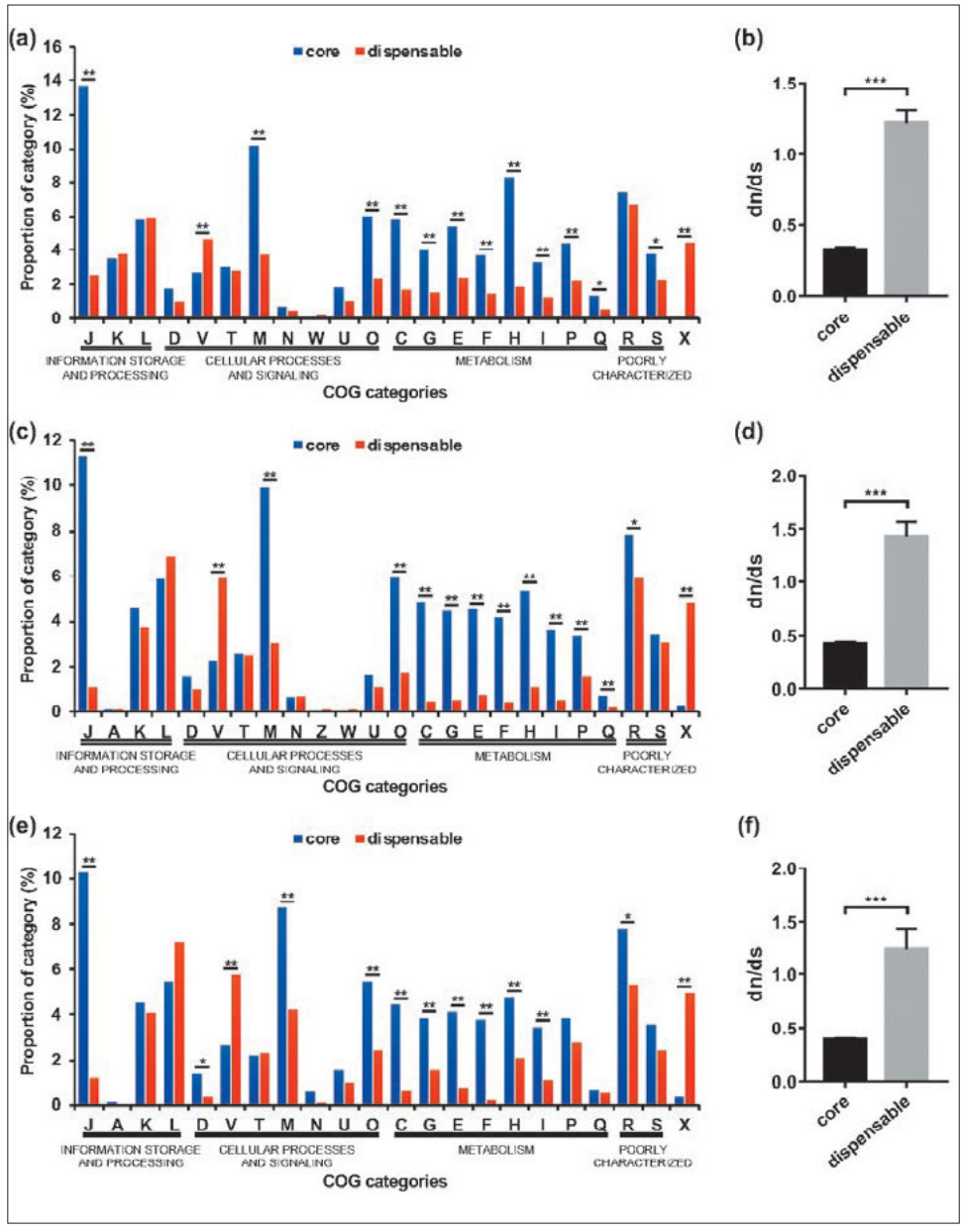

Fig 1 Functional analysis. Comparison of the abundance for each COG category and dn/ds values of homogenous clusters in the core and dispensable genome in *P. gingivalis* **(a and b)**, *P. intermedia* **(c and d)** and *P. nigrescens* **(e and f)**. The asterisks indicate those that are significantly different (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). For all three species, the enriched genes in the core genome were involved in almost all the COG categories of metabolism. The dispensable genome was enriched in the genes related to defence (category V) and mobilome (category X) **(a, c and e)**. The mean value of dn/ds for the core genome was below 1 in all three species, which was significantly lower than that of the dispensable genome **(b, d and f)**. See Supplementary Table S3 (provided on request) for COG codes.

Pan-genome and core genome analysis

The pan-genome and core genome development plots of *P. gingivalis*, *P. intermedia* and *P. nigrescens* are shown in Fig 2. All three species were found to possess an open pan-genome that continues to increase in the number of genes as new genomes are added (Figs 2a, 2c and 2e). On average, with each new genome sequenced, there were additions of 31, 68 and 145 genes to the pan-genome of *P. gingivalis*, *P. intermedia* and *P. nigrescens*, respectively.

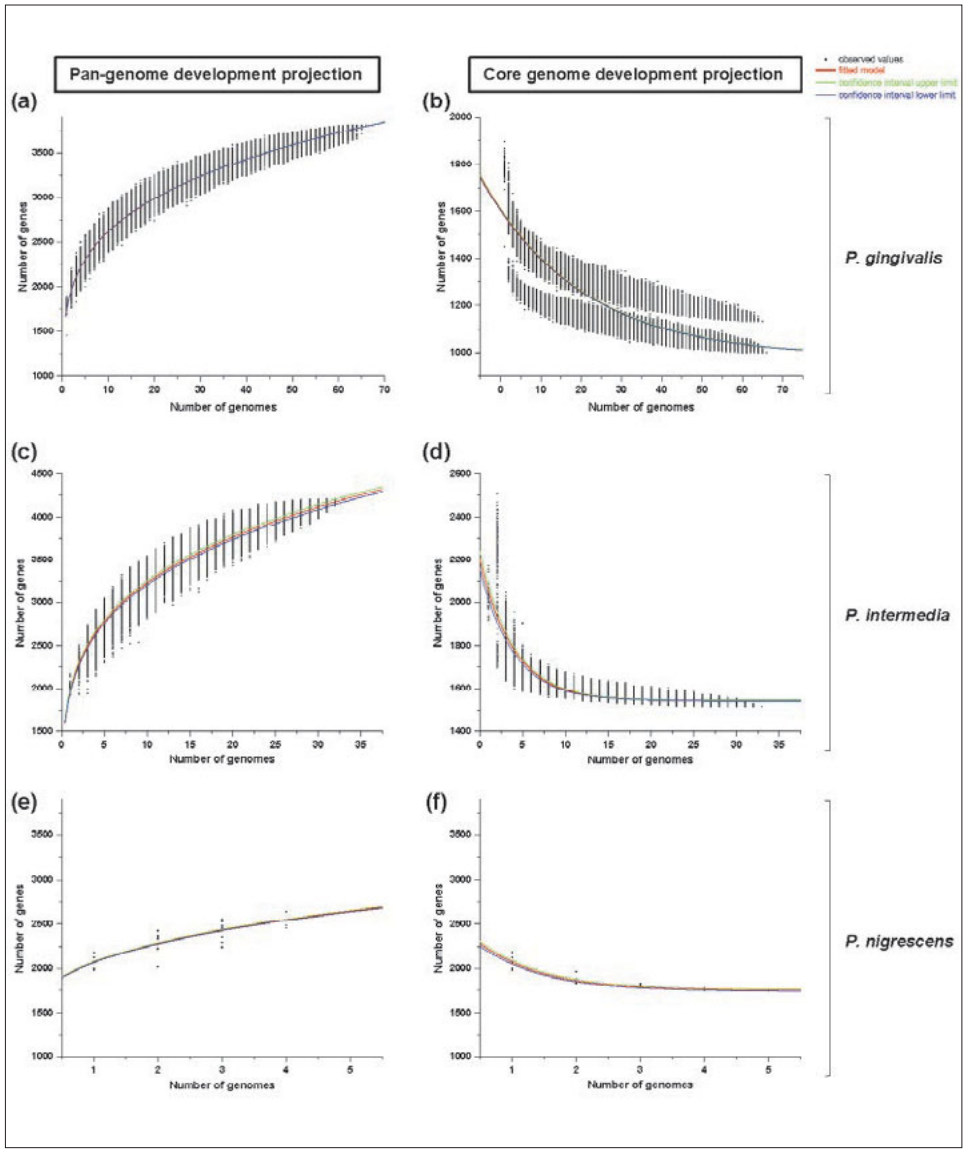

Fig 2 Pan-genome and core genome development plot projections for *P. gingivalis* **(a and b)**, *P. intermedia* **(c and d)** and *P. nigrescens* **(e and f)**. In the six panels, each black point represents the size of the pan-genome or core genome calculated from random combinations and iterations of strains as genomes are included in the analysis. For pan-genome development plot extrapolation: the red curve shows the fitted exponential Heaps' law function, and the green and blue curves indicate the upper and lower boundary of the 95% confidence interval. For all three species, the extrapolated curves continue to increase, indicating that they all have an open pan-genome. For core genome development plot extrapolation: the red curve shows the fitted exponential decay function, and the green and blue curves indicate the upper and lower boundary of the 95% confidence interval. The size of the core genome decreases at a slower rate with the addition of each strain genome, and remains relatively constant even as more genomes are added.

The significant difference should be due to the variation in the number of sequenced genomes. Since the number of *P. nigrescens* genomes is much smaller, it is anticipated that with the addition of further genomes, more new genes are likely to be added to the pool.

Aside from these black-pigmented species, an open pan-genome is also found in *streptococcus agalactiae*[6], *escherichia coli*[11], *streptococcus pneumoniae*[37] and *streptococcus mutans*[22]. These bacterial species exhibit high genome expansion plasticity through the lateral exchange of genetic material to adapt to variable environments. On the contrary, some bacteria display a closed pan-genome, such as *staphylococcus aureus*[41], *staphylococcus lugdunensis*[42], *salmonella enterica*[43] and *yersinia pestis*[44]. For these species, any acquisition of foreign genes seems to be largely restricted to bacteriophages and plasmids.

The core genome of the three species also displayed a similar evolution. The core genome size decreased at a slower rate with the addition of each new sequence, and then remained relatively constant even as more genomes were added (Figs 2b, 2d and 2f). The extrapolated core genome sizes were 980 (95% confidence interval 979–980 genes), 1543 (95% confidence interval 1542–1543 genes) and 1746 (95% confidence interval 1736–1756 genes) genes for *P. gingivalis*, *P. intermedia* and *P. nigrescens*, respectively (Supplementary data 1, Table S1, provided on request). This suggested that all three species are defined by a limited number of core genes and display a degree of variability in dispensable genes.

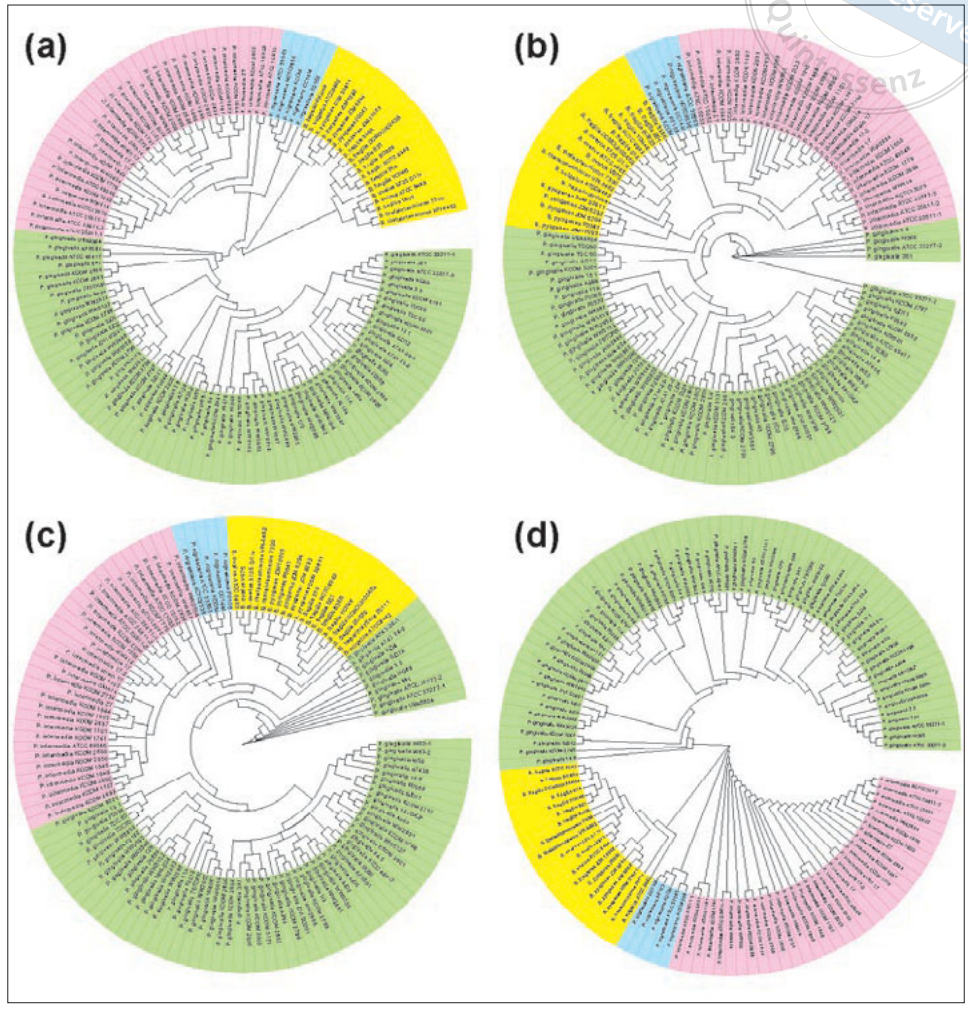

Fig 3 Circular phylogenetic trees illustrating the genetic relationships between strains of black-pigmented species. **(a)** The tree was built based on the pan-genome using the UPGMA method. **(b)** The tree was built based on the pan-genome using the neighbour-joining method. **(c)** The tree was built based on SNPs within the core genome using the maximum likelihood method. **(d)** The tree was built based on the 16S rRNA gene using the maximum likelihood method. For each tree, the leaf colour of light green, pink, light blue and yellow represents *P. gingivalis*, *P. intermedia*, *P. nigrescens* and species of *Bacteroides*, respectively. All the trees show that *P. gingivalis*, *P. intermedia*, *P. nigrescens* and species of *Bacteroides* form four clearly separated clusters within the phylogenetic trees, and there is a closer relationship between *P. intermedia* and *P. nigrescens*.

Phylogenetic analysis

The phylogenetic tree is a common methodology to infer the evolutionary relationship between different species. To confirm the close phylogenesis of *P. gingivalis*, *P. intermedia* and *P. nigrescens*, we constructed phylogenetic trees for the three species together based on the pan-genome profile (Figs 3a and b), SNP information for the core genome (Fig 3c) and 16S rRNA gene (Fig 3d). We compared these trees and found that although they were not completely identical, they exhibited considerably similar topology. All the trees showed a clear separation in *P. gingivalis*, *P. intermedia*, *P. nigrescens* and species of *Bacteroides*. They all formed monophyletic branches, and strains of the same species were closest to each other. It is worth noting that there is a close evolutionary relationship between *P. intermedia* and *P. nigrescens* as they originate from the same branch in all the phylogenetic trees. Among the four trees based on different genome information with diverse algorithms, the pan-genome tree using the unweighted pair group method with arithmetic mean (UPGMA) exhibited a higher resolution at the species level, with some already known closely related isolates located in the adjacent branches, such as *P. intermedia* ATCC 25611-1, *P. intermedia* ATCC 25611-2 and *P. intermedia* ATCC 25611-3. This verified not only the superiority of the UPGMA algorithm in constructing phylogenetic trees for closely related species, but also the advantage of the pan-genome in phylogenetic analysis, especially for species with incomplete or unavailable 16S rRNA gene sequences.

To reveal the influences of geographic locations on the evolution of *P. gingivalis*, *P. intermedia* and *P. nigrescens*, we referred to the NCBI websites and previously published studies for the geographic locations of each taxon and combined the geographic locations with the phylogenetic tree. The geographic locations of the analysed taxa included the USA, UK, South Korea, China, Japan, Sudan, Germany, Norway, Chile, Canada, Romania, Russia, Denmark and Hungary, although this

information was missing for a few taxa. The strains with different geographic locations are scattered throughout different branches on the phylogenetic trees, which suggested that adapting to different countries did not play an essential role in the evolutionary history of the black-pigmented bacteria.

Furthermore, the average nucleotide identity (ANI) value, which is a widely accepted genome-based method for species delineation, was calculated for all the strains and represented as a heatmap (Supplementary Figure, provided on request). The heatmap shows the degree of similarity between strains based on the ANI values of the whole-genome sequences. In agreement with phylogenetic analysis, the resulting ANI values of *P. gingivalis*, *P. intermedia* and *P. nigrescens* were distributed between 97.49% and 99.99%, 95.16% and 99.99%, 96.93% and 99.91% (Supplementary data 2, Excel S1, provide on request), respectively, meeting the cut-off criteria of 95% identity as suggested previously[29]. The ANI values between the three species and Bacteroides were distributed between 63.76% and 71.78% (Supplementary data 2, Excel S1), which is much lower than the cut-off of 95% similarity, further confirming the previously modified nomenclature for these closely related black-pigmented bacteria. It is worth noting that the ANI values of *P. intermedia* and *P. nigrescens* were distributed between 84.12% and 85.92% (Supplementary data 2, Excel S1), which was much higher than that of other different species, further validating their close evolutionary relationship.

Distribution of virulence factors

We reviewed the literature and found that *P. gingivalis*, *P. intermedia* and *P. nigrescens* possess a range of virulence factors involved in adhesion, proteolysis and evasion of host defences[2,32-34]. The three species share almost the same virulence factors, including capsule, lipopolysaccharide (LPS) and hemagglutinin[32-34]. Some virulence factors are conserved in all three species, such as rubrerythrin (Fig 4a), which is a non-heme iron protein that protects many air-sensitive bacteria against oxidative stress[45]. Genes coding for serine phosphatase (SerB), which has been shown to inhibit the biosynthesis of cytokines in gingival epithelial cells[33], were also conserved in *P. gingivalis*, *P. intermedia* and *P. nigrescens* (Fig 4a). For most virulence factors, the level of conservation varies with the aspect of pathogenicity. The involved genes have a relatively conserved part and a strain-specific part (Fig 4a).

To invade periodontal tissues, the periodontal pathogens need to adhere to the surface of the tooth or mucosa, which is facilitated by a repertoire of adhesins, including fimbriae and hemagglutinin[46]. Fimbriae are thin, filamentous, cell-surface protrusions that facilitate adherence of the periodontal pathogens to salivary proteins, extracellular matrix and bacteria of either the same or other species. Through the fimbriae, they can attach to the early colonising bacteria and participate in the developing biofilm structure. Hemagglutinin is a type of non-fimbrial adhesin that agglutinates erythrocytes and promotes adherence to host tissue, including endothelial cells. In *P. gingivalis* strains, some fimbriae and hemagglutinin genes are conserved, whereas others only exist in certain strains. In *P. nigrescens*, genes coding for fimbriae and hemagglutinin belong to the dispensable genome and the core genome, respectively (Fig 4a).

Apart from adhesins, LPS is also a key virulence factor in the development of periodontitis. LPS is a major component of the outer cell membrane of gram-negative bacteria, including periodontal pathogens *P. gingivalis*, *P. intermedia* and *P. nigrescens*. It is recognised by the host and potently activates host inflammatory and innate defence responses[34]. LPS has been demonstrated to be a stimulator of proinflammatory responses and bone resorption in animal models of periodontitis[47,48]. In *P. gingivalis*, both gene copies coding for LPS belong to the core genome. Notably, one of them has the conservation value of 65 and is only absent in *P. gingivalis* Ando strain (Fig 4b). Since the genome of *P. gingivalis* Ando is not completely assembled (scaffold level), some genes might have been missed during sequencing and assembly of the draft genome. Thus, this gene is still regarded as conserved. In *P. intermedia* and *P. nigrescens*, there are also two gene copies coding for LPS shared by all the strains, while other genes are only found in one or several strains (Supplementary data 3, Excel S2, provided on request).

Like LPS, the capsule is an outer envelope structure that lies outside bacterial cells. Many bacterial strains have been found to possess capsules, such as *Streptococcus pneumonia*, whose capsule was verified as an important virulence factor in the early 20th century[49]. Similar to its functions in other bacteria, the capsule of *P. gingivalis* protects the bacteria from host immune clearance. It has been reported that encapsulated strains of *P. gingivalis* are more resistant to phagocytosis by polymorphonuclear leukocytes than non-encapsulated strains[50]. In this study, all three genes coding for capsule biosynthesis proteins are conserved in *P. gingivalis*, whereas in *P. intermedia* and *P. nigrescens*, genes coding for the capsule belong to the dispensable genome. Interestingly, higher conservation

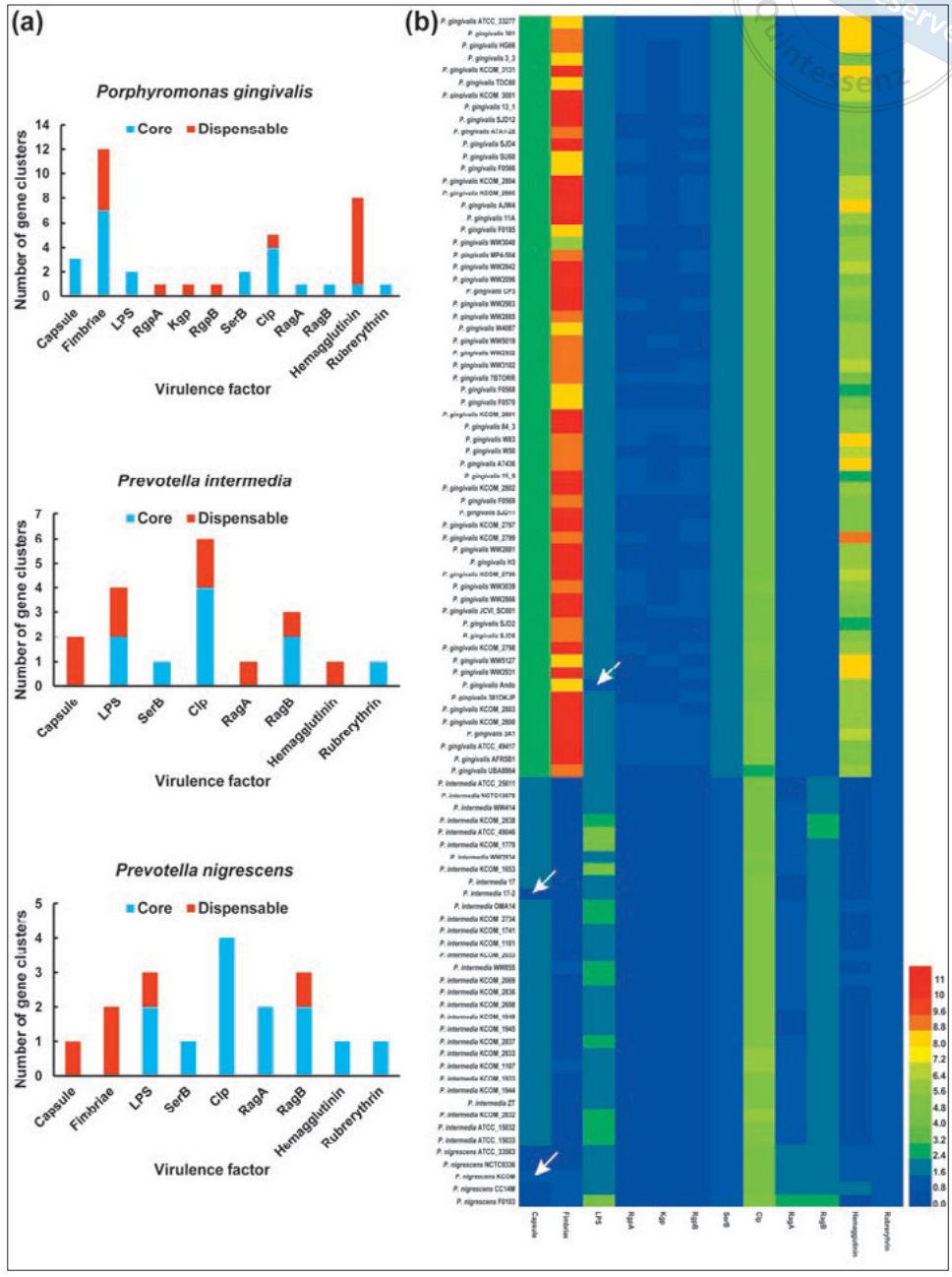

Fig 4 Distribution of virulence factors. **(a)** The overall number of putative virulence genes and their core and dispensable fractions are indicated for each virulence gene. Both SerB and rubrerythrin are conserved in all three species, whereas for most virulence factors, there is a conserved part and a dispensable part. **(b)** Heatmap illustrating the distribution and abundance of putative virulence genes across important black-pigmented strains. The gene copy number of each virulence gene is indicated by the colour key ranging from blue (absent) to red. Strains are graphed top to bottom in the same order as they appear in the phylogeny based on the pan-genome using UPGMA (Fig 3a). The white arrows highlight the representative strains. In *P. intermedia*, genes coding for the capsule show a high conservation level. Both of the two gene copies are only absent in *P. intermedia* 17-2, whereas in *P. nigrescens*, the single gene coding for the capsule is specific to *P. nigrescens* KCOM. In *P. gingivalis*, one of the genes coding for LPS is only absent in *P. gingivalis* Ando, whose genome is not completely assembled (scaffold level). Thus, we still regard this gene as conserved.

levels of the capsule were found in *P. intermedia*. Both gene copies coding for the capsule are only absent in *P. intermedia* 17-2 strain (Fig 4b), which is a complete genome, whereas in *P. nigrescens*, the single gene coding for the capsule is specific to *P. nigrescens* KCOM (Fig 4b). It seems that both genes coding for LPS and capsule were conserved in *P. gingivalis* but showed lower conservation levels in *P. intermedia* and *P. nigrescens*, which is consistent with previous studies reporting that LPS and capsule are only found in certain strains of *P. intermedia* and *P. nigrescens*, especially those isolated from periodontitis patients[15].

Another major virulence factor in black-pigmented periodontal pathogens are proteases. The most important proteases in *P. gingivalis* are gingipains, which account for 85% of the extracellular proteolytic activities of *P. gingivalis*[51]. Aside from their role in degrading host tissue, gingipains are also involved in enhancing the interactions of *P. gingivalis* with other periodontal pathogens to facilitate bacterial survival and biofilm formation[52]. Based on their specificity, gingipains can be divided into two groups: arginine-specific (Rgp) and lysine-specific (Kgp). Rgp is further subdivided into RgpA and RgpB based on structure. All the genes coding

for RgpA, RgpB and Kgp are single-copy and belong to the dispensable genome (Figs 4a and b). Interestingly, genes coding for RgpA and Kgp are attributed to the same homogenous cluster in this study (Supplementary data 3, Excel S2, provided on request), indicating their sequence similarity. This is consistent with a previous report that found RgpA and Kgp have a similar structure and close molecular weights[35], whereas RgpB has a smaller molecular weight and lacks the C-terminal hemagglutinin domain compared with RgpA and Kgp[53].

Although gingipains are specific to *P. gingivalis*, another proteolytic enzyme belonging to the caseinolytic protease (Clp) family is ubiquitous among various organisms[54]. The Clp family has several members, including ClpB, ClpC, ClpP and ClpX, all of which play an important role in colonisation and survival in the oral cavity. It has been reported that ClpC, ClpP and ClpX are necessary for *P. gingivalis* to enter host epithelial cells, and the absence of these proteases can result in diminished tolerance to high temperature in *P. gingivalis*. ClpB does not play a role in entry, but is required for intracellular replication and survival in *P. gingivalis*[55]. In this study, all three species had an intact repertoire of the four gene copies coding for ClpB, ClpC, ClpP and ClpX in the core genome (Fig 4a and Supplementary data 3, Excel S2), which is consistent with a previous study that reported that Clp proteases were highly conserved[54], indicating the importance of the Clp family in *P. gingivalis*, *P. intermedia* and *P. nigrescens*.

Conclusion

Pan-genome analysis at the intraspecies level of three clinically important, black-pigmented periodontal pathogens was performed. We identified an open pan-genome structure for all three species. Phylogenic analysis presented a clear separation of *P. gingivalis*, *P. intermedia*, *P. nigrescens* and species of *Bacteroides*, verifying the reclassification of these black-pigmented species. All three species were found to share almost the same virulence factors involved in adhesion, proteolysis and evasion of host defences. Genes coding for these virulence factors are flexible, displaying both strain specificity and universality. Various conservation levels of virulence genes indicated that the different strains of *P. gingivalis*, *P. intermedia* and *P. nigrescens* inherited a basic package of genes that enables them to adapt to the complicated oral environment and cause disease; however, this basic gene package diverged under different conditions and the genes of different functions evolved at different rates. The dataset setup in this study can act as a powerful framework for the addition of further sequenced strains, enabling the refinement of the pan-genome. The tools and approaches used could also be applied as a pan-genomics framework to other species in future studies.

Conflicts of interest

The authors declare no conflicts of interest related to this study.

Author contribution

All authors contributed to the study conception and design. Dr Pei Qi MENG contributed to the functional annotation and phylogenetic analysis and drafted the manuscript; Dr Qian ZHANG contributed to the data collection and analysis. All authors approved the final version of the manuscript.

(Received Jun 28, 2022; accepted Mar 2, 2023)

References

1. Dahlén GG. Black-pigmented gram-negative anaerobes in periodontitis. FEMS Immunol Med Microbiol 1993;6:181–192.
2. Lang NP, Lindhe J. Clinical Periodontology and Implant Dentistry, Two-Volume Set. Chichester: John Wiley & Sons, 2015.
3. Shah HN, Gharbia SE. Biochemical and chemical studies on strains designated Prevotella intermedia and proposal of a new pigmented species, Prevotella nigrescens sp. nov. Int J Syst Bacteriol 1992;42:542–546.
4. Nelson KE, Fleischmann RD, DeBoy RT, et al. Complete genome sequence of the oral pathogenic Bacterium porphyromonas gingivalis strain W83. J Bacteriol 2003;185:5591–5601.
5. Nambu T, Yamane K, Maruyama H, Mashimo C, Yamanaka T. Complete genome sequence of Prevotella intermedia strain 17-2. Genome Announc 2015;3:e00951-15.
6. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 2005;102:13950–13955.
7. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005;15:589–594.
8. Sun Z, Harris HM, McCann A, et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun 2015;6:8322.
9. Inglin RC, Meile L, Stevens MJA. Clustering of pan- and core-genome of Lactobacillus provides novel evolutionary insights for differentiation. BMC Genomics 2018;19:284.
10. Udaondo Z, Duque E, Ramos JL. The pangenome of the genus Clostridium. Environ Microbiol 2017;19:2588–2603.
11. Rasko DA, Rosovitz MJ, Myers GS, et al. The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 2008;190:6881–6893.

12. Mosquera-Rendón J, Rada-Bravo AM, Cárdenas-Brito S, Corredor M, Restrepo-Pineda E, Benítez-Páez A. Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species. BMC Genomics 2016;17:45.
13. Chen T, Siddiqui H, Olsen I. In silico comparison of 19 Porphyromonas gingivalis strains in genomics, phylogenetics, phylogenomics and functional genomics. Front Cell Infect Microbiol 2017;7:28.
14. Chen T, Siddiqui H, Olsen I. Comparative genomics and proteomics of 13 Porphyromonas gingivalis strains. J Oral Microbiol 2015;7:29008.
15. Zhang Y, Zhen M, Zhan Y, Song Y, Zhang Q, Wang J. Population-genomic insights into variation in Prevotella intermedia and Prevotella nigrescens isolates and its association with periodontal disease. Front Cell Infect Microbiol 2017;7:409.
16. Ruan Y, Shen L, Zou Y, et al. Comparative genome analysis of Prevotella intermedia strain isolated from infected root canal reveals features related to pathogenicity and adaptation. BMC Genomics 2015;16:122.
17. Fischer S, Brunk BP, Chen F, et al. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics 2011;35:6.12.1-6.12.19.
18. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 2008;24:319–324.
19. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30:1575–1584.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
21. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: Pan-genomes analysis pipeline. Bioinformatics 2012;28:416–418.
22. Meng P, Lu C, Zhang Q, Lin J, Chen F. Exploring the genomic diversity and cariogenic differences of Streptococcus mutans strains through pan-genome and comparative genome analysis. Curr Microbiol 2017;74:1200–1209.
23. Song L, Wang W, Conrads G, et al. Genetic variability of mutans streptococci revealed by wide whole-genome sequencing. BMC Genomics 2013;14:430.
24. Heaps HS. Information Retrieval: Computational and Theoretical Aspects. New York: Academic Press, 1978.
25. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797.
26. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35:3100–3108.
27. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. Mol Biol Evol 2016;33:1870–1874.
28. Zhang H, Gao S, Lercher MJ, Hu S, Chen WH. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. Nucleic Acids Res 2012;40:W569–W572.
29. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 2009;106:19126–19131.
30. Deng W, Wang Y, Liu Z, Cheng H, Xue Y. HemI: A toolkit for illustrating heatmaps. PLoS One 2014;9:e111988.
31. Okamoto M, Maeda N, Kondo K, Leung KP. Hemolytic and hemagglutinating activities of Prevotella intermedia and Prevotella nigrescens. FEMS Microbiol Lett 1999;178:299–304.
32. Zenobia C, Hajishengallis G. Porphyromonas gingivalis virulence factors involved in subversion of leukocytes and microbial dysbiosis. Virulence 2015;6:236–243.
33. Sochalska M, Potempa J. Manipulation of neutrophils by Porphyromonas gingivalis in the development of periodontitis. Front Cell Infect Microbiol 2017;7:197.
34. Mysak J, Podzimek S, Sommerova P, et al. Porphyromonas gingivalis: Major periodontopathic pathogen overview. J Immunol Res 2014;2014:476068.
35. Jia L, Han N, Du J, Guo L, Luo Z, Liu Y. Pathogenesis of important virulence factors of Porphyromonas gingivalis via toll-like receptors. Front Cell Infect Microbiol 2019;9:262.
36. Xu W, Zhou W, Wang H, Liang S. Roles of Porphyromonas gingivalis and its virulence factors in periodontitis. Adv Protein Chem Struct Biol 2020;120:45–84.
37. Park SC, Lee K, Kim YO, Won S, Chun J. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. Front Microbiol 2019;10:834.
38. Galardini M, Mengoni A, Brilli M, et al. Exploring the symbiotic pangenome of the nitrogen-fixing bacterium Sinorhizobium meliloti. BMC Genomics 2011;12:235.
39. Bottacini F, Medini D, Pavesi A, et al. Comparative genomics of the genus Bifidobacterium. Microbiology (Reading) 2010;156:3243–3254.
40. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet 2008;4:e1000304.
41. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science 2010;327:469–474.
42. Argemi X, Matelska D, Ginalski K, et al. Comparative genomic analysis of Staphylococcus lugdunensis shows a closed pan-genome and multiple barriers to horizontal gene transfer. BMC Genomics 2018;19:621.
43. Holt KE, Parkhill J, Mazzoni CJ, et al. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat Genet 2008;40:987–993.
44. Cui Y, Yu C, Yan Y, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci U S A 2013;110:577–582.
45. Sztukowska M, Bugno M, Potempa J, Travis J, Kurtz DM Jr. Role of rubrerythrin in the oxidative stress response of Porphyromonas gingivalis. Mol Microbiol 2002;44:479–488.
46. Lamont RJ, Jenkinson HF. Subgingival colonization by Porphyromonas gingivalis. Oral Microbiol Immunol 2000;15:341–349.
47. Chiang CY, Kyritsis G, Graves DT, Amar S. Interleukin-1 and tumor necrosis factor activities partially account for calvarial bone resorption induced by local injection of lipopolysaccharide. Infect Immun 1999;67:4231–4236.
48. Nishida E, Hara Y, Kaneko T, Ikeda Y, Ukai T, Kato I. Bone resorption and local interleukin-1alpha and interleukin-1beta synthesis induced by Actinobacillus actinomycetemcomitans and Porphyromonas gingivalis lipopolysaccharide. J Periodontal Res 2001;36:1–8.
49. Griffith F. The significance of pneumococcal types. J Hyg (Lond) 1928;27:113–159.
50. Sundqvist G, Figdor D, Hänström L, Sörlin S, Sandström G. Phagocytosis and virulence of different strains of Porphyromonas gingivalis. Scand J Dent Res 1991;99:117–129.
51. de Diego I, Veillard F, Sztukowska MN, et al. Structure and mechanism of cysteine peptidase gingipain K (Kgp), a major virulence factor of Porphyromonas gingivalis in periodontitis. J Biol Chem 2014;289:32291–32302.
52. Bao K, Belibasakis GN, Thurnheer T, Aduse-Opoku J, Curtis MA, Bostanci N. Role of Porphyromonas gingivalis gingipains in multi-species biofilm formation. BMC Microbiol 2014;14:258.

53. Guo Y, Nguyen KA, Potempa J. Dichotomy of gingipains action as virulence factors: From cleaving substrates with the precision of a surgeon's knife to a meat chopper-like brutal degradation of proteins. Periodontol 2000 2010;54:15–44.
54. Kress W, Maglica Z, Weber-Ban E. Clp chaperone-proteases: Structure and function. Res Microbiol 2009;160:618–628.
55. Capestany CA, Tribble GD, Maeda K, Demuth DR, Lamont RJ. Role of the Clp system in stress tolerance, biofilm formation, and intracellular invasion in Porphyromonas gingivalis. J Bacteriol 2008;190:1436–1446.