


RESEARCH ARTICLE

Open Access

Comparative genomics reveals broad genetic diversity, extensive recombination and nascent ecological adaptation in *Micrococcus luteus*



Yisong Li, Zhong-Zhi Sun, Jin-Cheng Rong and Bin-Bin Xie* 

Abstract

Background: *Micrococcus luteus* is a group of actinobacteria that is widely used in biotechnology and is being thought as an emerging nosocomial pathogen. With one of the smallest genomes of free-living actinobacteria, it is found in a wide range of environments, but intraspecies genetic diversity and adaptation strategies to various environments remain unclear. Here, comparative genomics, phylogenomics, and genome-wide association studies were used to investigate the genomic diversity, evolutionary history, and the potential ecological differentiation of the species.

Results: High-quality genomes of 66 *M. luteus* strains were downloaded from the NCBI GenBank database and core and pan-genome analysis revealed a considerable intraspecies heterogeneity. Phylogenomic analysis, gene content comparison, and average nucleotide identity calculation consistently indicated that the species has diverged into three well-differentiated clades. Population structure analysis further suggested the existence of an unknown ancestor or the fourth, yet unsampled, clade. Reconstruction of gene gain/loss events along the evolutionary history revealed both early events that contributed to the inter-clade divergence and recent events leading to the intra-clade diversity. We also found convincing evidence that recombination has played a key role in the evolutionary process of the species, with upto two-thirds of the core genes having been affected by recombination. Furthermore, distribution of mammal-associated strains (including pathogens) on the phylogenetic tree suggested that the last common ancestor had a free-living lifestyle, and a few recently diverged lineages have developed a mammal-associated lifestyle separately. Consistently, genome-wide association analysis revealed that mammal-associated strains from different lineages shared genes functionally relevant to the host-associated lifestyle, indicating a recent ecological adaption to the new host-associated habitats.

(Continued on next page)

* Correspondence: xbb@sdu.edu.cn

Microbial Technology Institute and State Key Laboratory of Microbial Technology, Shandong University, Qingdao 266237, China



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

Conclusions: These results revealed high intraspecies genomic diversity of *M. luteus* and highlighted that gene gain/loss events and extensive recombination events played key roles in the genome evolution. Our study also indicated that, as a free-living species, some lineages have recently developed or are developing a mammal-associated lifestyle. This study provides insights into the mechanisms that drive the genome evolution and adaptation to various environments of a bacterial species.

Keywords: *Micrococcus luteus*, Population genomics, Pan-genome, Intraspecies diversity, Recombination, Adaptive evolution, Ecological differentiation

Background

Micrococcus luteus, the type species of the genus *Micrococcus* (family *Micrococcaceae*, order *Micrococcales*), is a high GC Gram-positive coccus of the phylum *Actinobacteria* [1]. *M. luteus* is known as an opportunistic pathogen for nosocomial infections [2], and has been proved to be able to cause bacteremia, pneumonia, endocarditis, lymphoma, septic arthritis and many other diseases [3–5]. Besides as a pathogen, *M. luteus* is ubiquitously distributed in a variety of habitats, including soil, air, marine, plant and the human body [6], indicating that this species has been well adapted to various environments. It was also found that *M. luteus* can resuscitate and stimulate ‘viable but non-culturable (VBNC)’ or uncultured bacteria, by secreting a small protein called resuscitation-promoting factor (Rpf) [7, 8], and this feature has already been used in many biotechnological fields [9–11]. Therefore, as an emerging nosocomial pathogen and a strain of biotechnological interest, *M. luteus* has received increasing attention in recent years. It was speculated that *M. luteus* is primarily adapted to mammalian skin, and that its occasional presence elsewhere, such as water or soil, might possibly arise from contamination by skin flakes [12]. However, the adaptation strategy to various environments and the underlying genetic basis remains largely unknown.

Comparative and population genomics have emerged as valuable tools to delimit species features and to explore mechanisms of environmental adaptation or even speciation. With the ever-increasing whole-genome sequencing of closely related populations of microorganisms, combined with simulations and modelling, it has been widely accepted that microbial speciation is usually driven by natural selection for adaptation to distinct ecological niches [13, 14]. During this process, genomic variation caused by horizontal transfer (gene gain), gene loss, and duplication, plays a significant role [15, 16]. For example, a population genomics study of vibrios revealed a large number of gene gain and loss events in their evolutionary history, enabling vibrios to occupy various niches [17]. Similarly, it was also shown that, in streptomycetes, genomic fluctuation could ensure a quick and economical response to various lifestyles [18, 19]. The

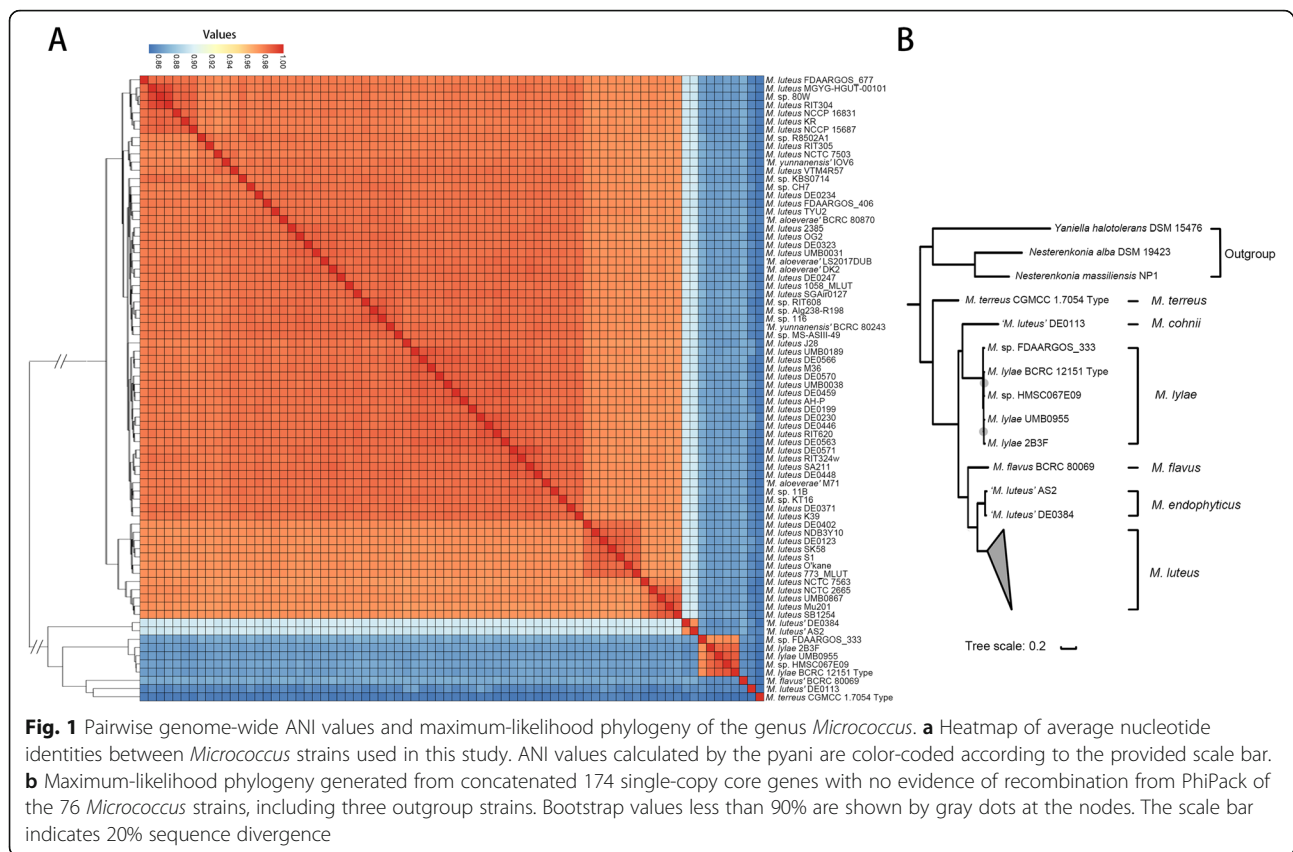
first sequenced genome of *M. luteus* revealed a circular chromosome with one of the smallest genomes of free-living actinobacteria and an abnormally high number of transposable elements [12]. However, the genomic diversity of *M. luteus* and the genomic events that contribute to the ecological adaptation need further study.

Now genome sequences of more than 70 *M. luteus* strains have been published and/or are available publicly [20–22], providing an opportunity to systematically study its physiology, ecology, and evolution at the genomic level. This study was aimed to investigate the intra-species genomic diversity of *M. luteus* and to uncover the evolutionary events that may contribute to the genomic diversity. Furthermore, this study was also intended to explore the relationships between the intra-species genomic diversity and the potential ecological differentiation within the species.

Results

Genomic features of *M. luteus*

All (106) *Micrococcus* genome sequences available at the NCBI (National Center for Biotechnology Information) assembly database (April 2020) were downloaded and subject to strict quality control to reduce potential bias in the subsequent analyses due to different assembly qualities (see Methods for details). As a result, 76 high-quality (completeness > 95% and contamination < 5%) and non-redundant *Micrococcus* genomes remained (Table S1). Whole-genome average nucleotide identity (ANI) analysis classified all these genomes into six clades (species) based on an ANI cutoff of 95% (Fig. 1a), three of which (including 72 strains) corresponded to three published species with inner ANI > 97.4%: *M. luteus* ($n = 66$), *M. lylae* ($n = 5$), and *M. terreus* ($n = 1$), respectively. As the other three clades did not contain type strain genomes, the remaining four strains (*‘Micrococcus luteus’* AS2, *‘Micrococcus luteus’* DE0384, *‘Micrococcus flavus’* BCRC 80069, and *‘Micrococcus luteus’* DE0113) in these three clades were then assigned to known species by comparing their near full-length 16S rRNA gene sequences to those of type strains. Strains AS2 and DE0384 were reclassified into *M. endophyticus* (type strain: YIM 56238^T, identity > 99.5%, 1432 nt)



[23], strain DE0113 was reclassified into *M. cohnii* (type strain: WS4601^T, identity = 100%, 1109 nt) [24], and strain BCRC 80069 was reclassified into *M. flavus* (type strain: LW4^T, identity = 99.9%, 1402 nt) [25]. Further phylogenomic analysis supported the species classification (Fig. 1b).

Our dataset contained genomes for 66 *M. luteus* strains. These strains were from diverse habitats globally, including marine, soil, food, plant- and human-associated, and other common sources (Table S1). Compared to high intra-species ANI values (> 96.8%), the ANI values with other species were much lower (< 89.4%), suggesting a clear inter-species genomic difference. The *M. luteus* genome size ranged from 2.41 Mb (strain OG2) to 2.74 Mb (strain DE0230) (Figure S1A), with an average genome size (2.52 Mb, $n = 66$) smaller than *M. lylae* (2.68 Mb, $n = 5$; $p = 0.001$, Wilcoxon test), *M. endophyticus* (2.78 Mb, $n = 2$; $p = 0.02$, Wilcoxon test), and *M. terreus* (3.09 Mb, $n = 1$), but larger than *M. cohnii* (2.24 Mb, $n = 1$). In addition, the GC content of *M. luteus* (mean 72.9%, $n = 66$; ranging from 72.3 to 73.3%) was higher than *M. lylae* (mean 71.3%, $n = 5$; $p = 0.0002$, Wilcoxon test), *M. cohnii* (70.8%, $n = 1$) and *M. terreus* (68.9%, $n = 1$), but lower than *M. endophyticus* (mean 73.4%, $n = 2$; $p = 0.02$, Wilcoxon test) and *M.*

flavus (73.5%, $n = 1$) (Figure S1B). Taken together, the differences in ANI values, genome size, and GC content, suggested an apparent divergence between *M. luteus* and the other species of *Micrococcus*.

The *M. luteus* genomes each had 2148 (strain OG2) to 2501 (strain SGAir0127) coding sequences (CDSs) (mean 2287). On average, 94.6% of them could be annotated and functionally categorized using the eggNOG database [26]. Each genome contained 159 to 315 identifiable insertion sequences (ISs), accounting for 7.22 to 13.4% of the genome (mean 8.51%). Each genome also harbored 4 to 15 genomic islands (GIs; mean 8.3) that were 3.1 to 96.2 kbp in size, which together accounted for an average of 6.6% of the whole genome (ranging from 2.3 to 11.4%). These results suggested that strains of *M. luteus* might have undergone frequent genomic exchanges [27]. In addition, 57 strains (86.4%) encoded Rpf, a secreted protein with an N-terminal transglycosylase-like domain (PF06737) and a C-terminal LysM domain (PF01476), which could stimulate the growth and resuscitation of dormant bacteria [7, 8]. This dormancy-resuscitation mechanism may help *M. luteus* to survive over extended periods when conditions are not conducive for growth, and to rapidly respond to environmental changes.

Potential virulence factors and antibiotic resistance genes of *M. luteus*

All *M. luteus* genomes were locally compared against the Virulence Factors Database (VFDB) [28] to detect virulence genes (Figure S2). We found 30 different putative virulence factors (VFs), twelve of which (40%) were related to GIs. Each genome contained 19 (strains RIT608 and 1058_MLUT) to 31 (strain SK58) VF genes (mean 22). Twelve VFs were shared by all strains, of which three (*clpC*, *clpP*, *katA*) were involved in stress response, three (*ideR*, *phoP*, *relA*) in regulation, two (*lirB*, *CBU_1566*) in secretion systems, and the remaining four (*htpB*, *gnd*, *bauE*, *icl*) in adherence, immune evasion, iron uptake and metabolic adaptation, respectively. These core VFs might play important roles in the pathogenicity of *M. luteus*. Copy number variation of the VF genes was also found. For example, genes *ureABG* were enriched (three to six copies) in strain SK58 (isolated from human skin). These genes are involved in bacteria urea hydrolysis and have been reported to be crucial to pathogenic bacteria virulence and defense against host immunity [29, 30]. Genes *msrA/B* (*pilB*) were expanded (three copies) in five strains and their gene products can promote the successful infection of humans and also respond to adverse conditions [31, 32]. The expansion of VF genes may promote the pathogenicity of the harboring strains.

We also detected the presence of antibiotic resistance genes (ARGs) in *M. luteus*. A total of 22 distinctive putative ARGs were identified, half of which were related to GIs (Figure S3). Overall, three ARGs were identified in all strains, including genes associated with macrolide- and penam-resistance (*mtrA*, with two copies), fosfomycin-resistance (*murA*) and rifamycin-resistance (*rbpA*). Additionally, an aminoglycoside phosphotransferase encoded by *strA*, which confers aminoglycoside-resistance, was found in 59.1% ($n = 39$) of the strains. The remaining genes ($n = 18$, 81.8%) exhibited sporadic distribution patterns (including eight strain-specific genes). These genes included 13 antibiotic efflux pump-encoding genes, three antibiotic inactivation enzymes encoding genes (*aac (3)-IIb*, *blaCTX-M-141*, *blmS*), and two genes for antibiotic target alteration (*rmtF*) and protection (*msrE*). Different ARG repertoire suggested that different strains may have different antibiotic resistance.

Core- and pan-genome of *M. luteus*

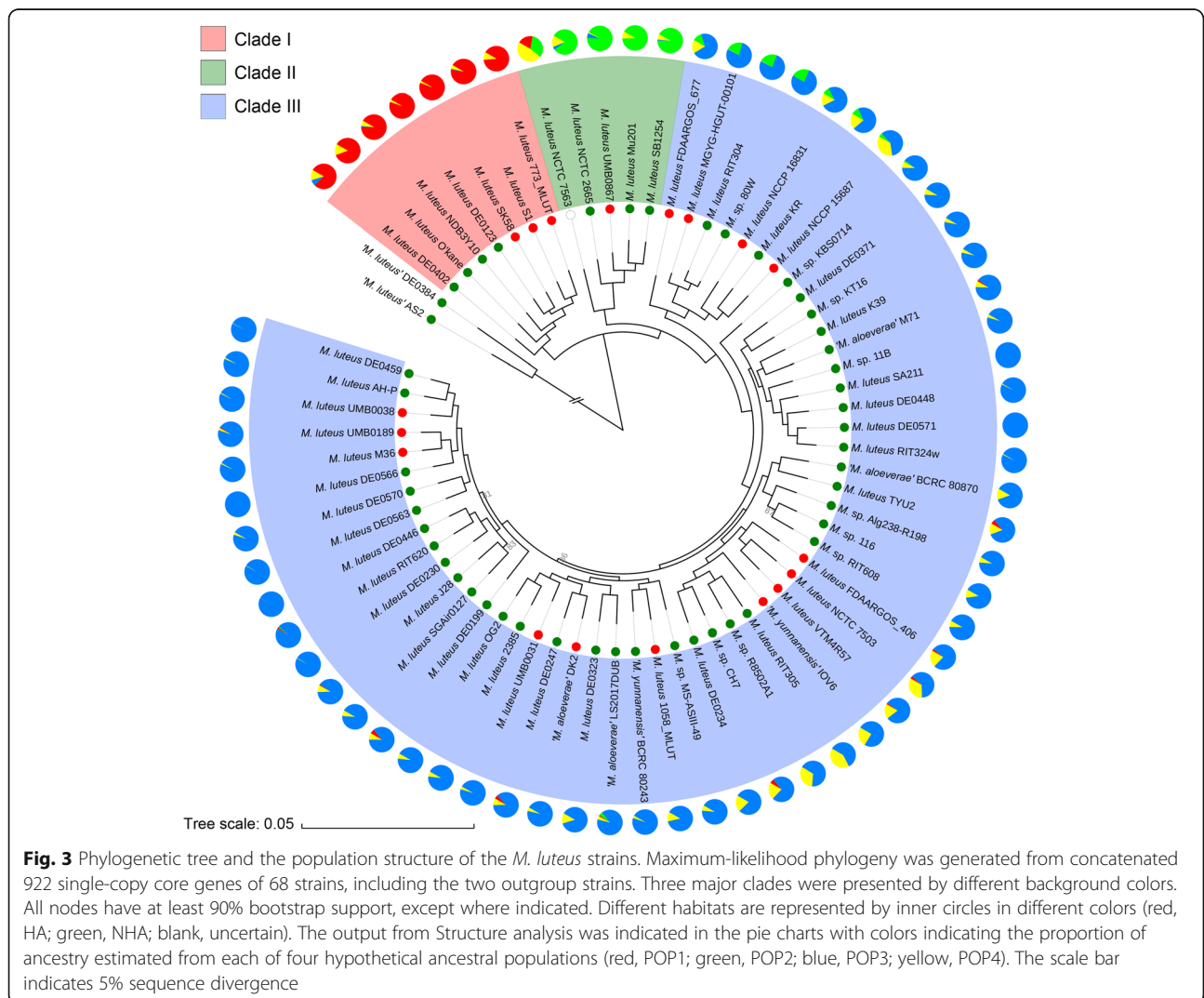
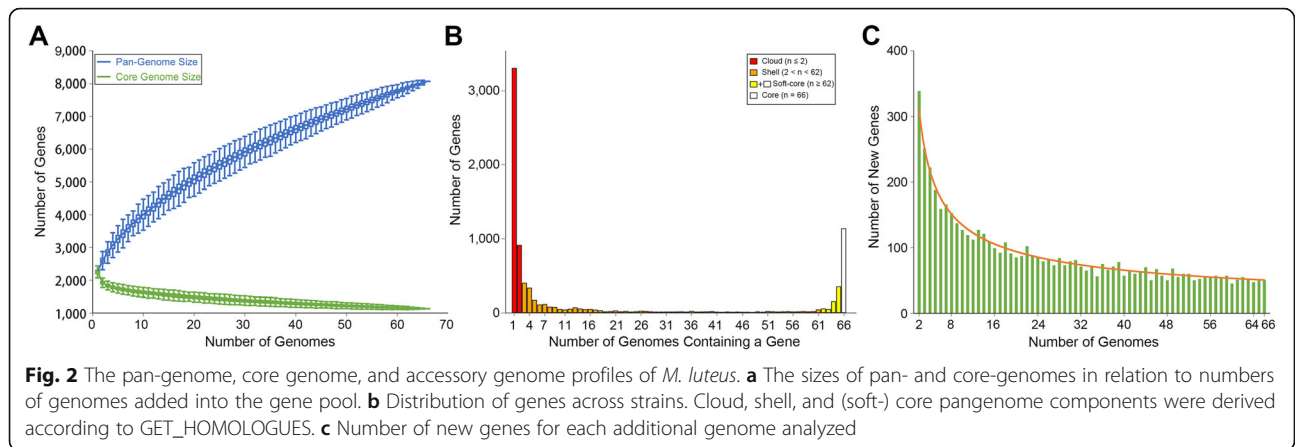
To explore the entire genomic repertoire of the *M. luteus* population, estimates of the cloud (genes present in only one or two genomes), shell (genes present in 3–61 genomes), softcore (genes present in at least 62 genomes) and core (genes present in all 66 genomes) genomes were generated using GET_HOMOLOGUES

[33]. The 66 *M. luteus* strains had a pan-genome of 8077 genes and a core genome of 1134 (14.0%) genes, including 991 single-copy core genes (Fig. 2). The core genome only represented 45.3 to 52.8% of the gene content of each strain, illustrating a relatively high degree of genomic diversity. This core gene ratio was much lower than that of the actinobacterial species *Streptomyces albidoflavus* (65.3 to 73.0%, recalculated with the same clustering algorithm and parameters, intraspecies ANI > 98.1%) [18]. Furthermore, the cloud genes contained more than half of the pan-genome (4210 genes, 52.1%; 3.36% for each strain, on average), of which 3301 were strain-specific genes (singletons, 40.9%; 2.14% for each strain, on average), indicating an exceptionally flexible genome of *M. luteus*. Correspondingly, the pan-genome for *M. luteus* still increased with approximately 50 new genes after addition of a 66th genome (Fig. 2c). Analysis of the pan-genome curve using a power-law regression model confirmed that the pan-genome was open ($B_{\text{pan}} = 0.5$), as the curve did not approach a constant as more genomes were selected.

We next performed a Clusters of Orthologous Groups (COG) functional classification for each orthologous group (OG) to define possible differences in the functions encoded by the soft-core, cloud and shell genomes of *M. luteus* (Figure S4). As a result, the soft-core genome had a higher proportion of genes classified in COG categories C (energy production and conversion), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), H (coenzyme transport and metabolism) and J (translation, ribosomal structure and biogenesis) ($p < 0.01$, Fisher's exact test), all associated with basic biological functions. The cloud and shell genes were biased toward COG categories L (replication, recombination and repair) and V (defense mechanisms) ($p < 0.01$, Fisher's exact test), which may contribute to the intraspecies heterogeneity of the species, as these genes have been proved to play important roles in the acquisition of foreign DNA [34].

Phylogenomic relationship and population structure of *M. luteus*

To uncover the evolutionary process that led to the genomic diversity, we constructed a highly robust phylogenomic maximum-likelihood (ML) tree of the 66 *M. luteus* strains based on concatenated 922 single-copy core genes, using the two *M. endophyticus* strains ('*Micrococcus luteus*' AS2 and '*Micrococcus luteus*' DE0384) as outgroups (Fig. 3). Based on the phylogenomic tree, *M. luteus* could be divided into three basal clades: Clade I, II, and III. Further analyses revealed clear inter-clade genomic boundaries. Firstly, the ANI heatmap and clustering results (Fig. 1a) suggested that the species could be divided into three groups, with high



intra-group ANI and relatively low inter-group ANI. Comparison of this clustering result with the phylogenomic tree indicated that the three clades corresponded nicely with the three ANI groups, indicating the existence of inter-clade ANI boundaries. Secondly, in order to investigate whether there were gene content differences between the three clades, a hierarchical clustering tree based on the content of dispensable genes was constructed (Figure S5). This tree clearly showed three clades, which were the same as those in the core genome tree, indicating the existence of inter-clade gene content boundaries. These results suggested that the separation between the three clades has emerged, both in the core and accessory genes.

The intraspecies difference was also studied using methods which are widely used in population structure analyses. By using Fastbaps [35], the entire *M. luteus* population could be partitioned into three subpopulations at Level 2 (Figure S6). These three subpopulations corresponded to the three major clades on the phylogenomic tree. At Level 3 or higher, one additional subpopulation emerged, which corresponded to only strain NCTC 7563 in Clade II. Furthermore, an admixture model implemented in STRUCTURE [36] software was applied, which showed maximal posterior probability at $K = 4$, indicating the existence of four ancestral subpopulations (Fig. 3 and Figure S6). Among them, three ancestral subpopulations (POP1, POP2, and POP3) were mainly represented by the three major clades. It was noted that, the fourth ancestral subpopulation (POP4, yellow) existed in high proportions in all the three clades, with the highest proportion in the fourth subpopulation uncovered by the above Fastbaps analysis (i.e., strain NCTC 7563 in Clade II). However, since POP4 accounted for only 47.5% of the genome of NCTC 7563, it is unclear whether POP4 can be represented by the branch of strain NCTC 7563. It is possible that POP4 corresponded to an unsampled clade.

Gain and loss of genes during the evolution of *M. luteus*

To unravel the evolutionary events that contributed to the intraspecies gene content differences, gene contents of all ancestral nodes on the phylogenomic tree were predicted and the numbers of gene gain and loss events that have occurred on all branches were calculated using a parsimony method (Figure S7). As a result, the last common ancestor of *M. luteus* was inferred to possess 2028 gene families. The gene family numbers had only slightly increased during evolution, as the gene gain events were a bit more numerous than the gene loss events in most stages during the evolutionary process. Also, a relatively high number of gene content variances occurred at the divergence of the three major clades (77, 64 and 55 for Clades I to III, respectively), consistent

with inter-clade gene content differences mentioned above (Figure S5). Interestingly, 88.3% of total gene gain events and 82.8% of total gene loss events occurred at terminal nodes, suggesting that the pan-genome diversity of *M. luteus* is largely due to frequent but recent strain-specific gene gain/loss events.

High level of homologous recombination in *M. luteus* species

Recombination, especially homologous recombination, is one of the main forces shaping bacterial evolution [37, 38]. By recombination, bacteria can receive DNA fragments from both the same and other species and integrate them into their chromosomes. Here, a series of analyses were used to evaluate the extent of homologous recombination in the core genome of *M. luteus*. Firstly, a NeighborNet network [39] based on the concatenated single-copy core genes showed a reticulate structure, indicating a high level of non-vertical inheritance in phylogenies (Fig. 4a). Meanwhile, the pairwise homoplasy index (PHI) [40] statistic provided highly significant evidence for recombination within *M. luteus* species ($p = 0.00$).

We also used mcorr [41] to calculate the probability that a pair of genomes differs at one locus conditional on having differences at the other locus. The resulting correlation profile exhibited a monotonic decay (Fig. 4b), indicating the presence of recombination. Similar decaying correlation profiles have also been shown in other recombining bacteria, such as *Helicobacter pylori*, *Pseudomonas aeruginosa*, *Salmonella enterica* and *Cronobacter sakazakii* [41–43]. Meanwhile, mcorr showed the mean fragment size (f) of a recombination event in *M. luteus* species was 874.45 bp; the recombination coverage (c), which indicates the proportion of sites in the genome whose diversity has come from recombination events since its last common ancestor, was 0.46; and the ratio γ/μ , the relative rate of recombination to mutation, was 10.87 (Fig. 4c). All these recombination parameters above further confirmed a high recombination rate within *M. luteus*. A summary of the parameters of *M. luteus* and other typical species reported was shown in Table 1.

We next sought to identify the frequently recombining genes. Among the 991 single-copy core genes, 628 genes (63.37%) showed significant evidence of genetic recombination, with P -values (computed from 1000 permutations) lower than 0.05 in at least two of the three methods implemented in PhiPack [40]. This result was further confirmed by fastGEAR [45], by which a total of 708 genes (71.44%) were detected to have been affected by recent or ancestral recombination events (Fig. 4d). The most frequently recombining genes included *pabB*, *cstA*, *betT*, and *comEC*: *pabB* encodes an enzyme that

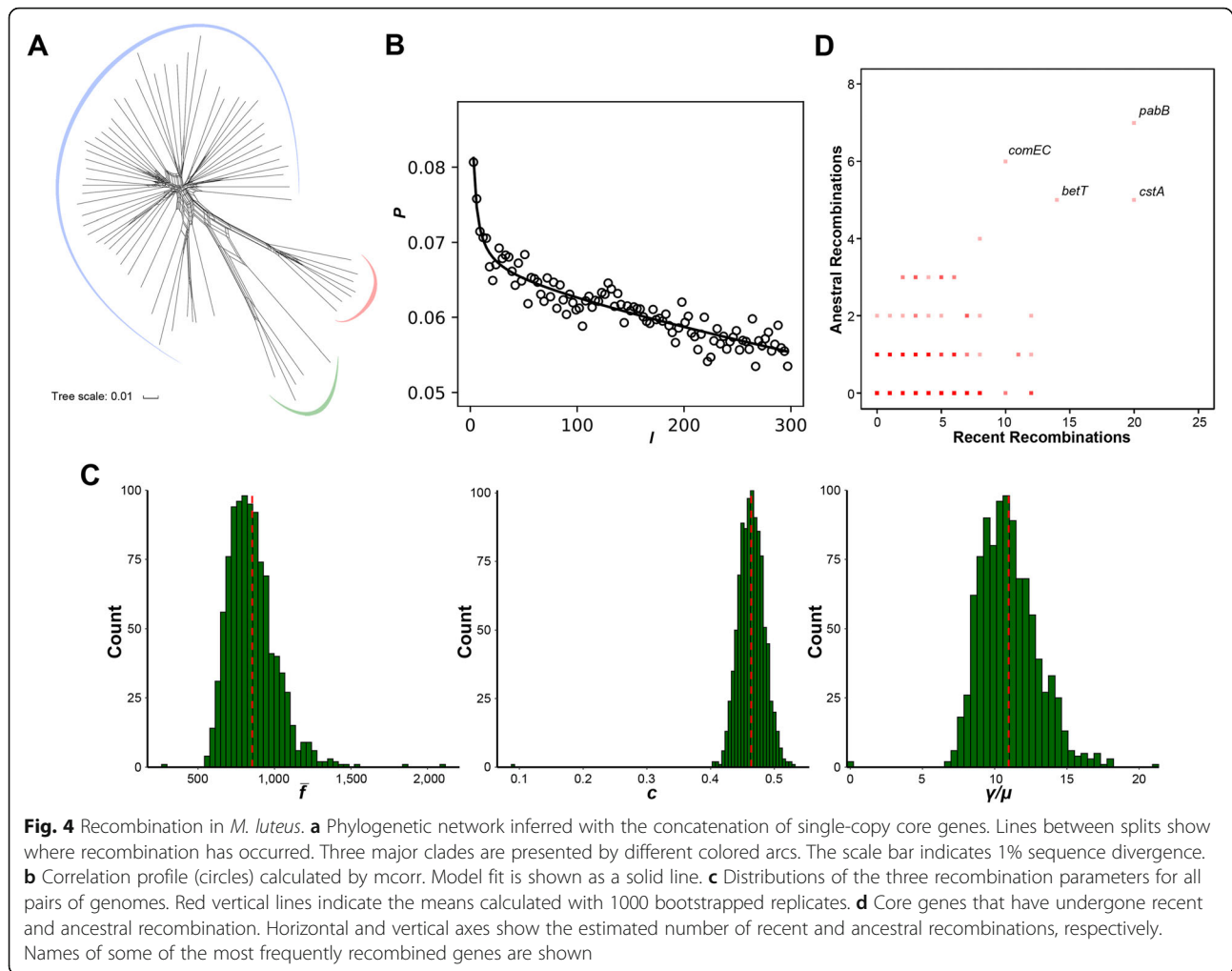


Table 1 Recombination parameters of different species inferred by mcorr

Species	\bar{f}	c	γ/μ	References
<i>Acinetobacter baumannii</i>	860	0.40	1.30	[41]
<i>Campylobacter jejuni</i>	1000	0.32	3.40	[41]
<i>Cronobacter sakazakii</i>	816	0.53	1.61	[43]
<i>Klebsiella pneumoniae</i>	5800	0.27	4.20	[41]
<i>Micrococcus luteus</i>	874	0.46	10.87	This study
<i>Micrococcus luteus</i> (NHA only)	903	0.45	10.41	This study
<i>Micrococcus luteus</i> (HA only)	849	0.49	10.93	This study
<i>Mycobacterium abscessus</i>	1200	0.21	13.00	[41]
<i>Pseudomonas aeruginosa</i>	590	0.52	11.00	[41]
<i>Salmonella enterica</i>	-	0.46	6.16	[42]
<i>Staphylococcus aureus</i>	550	0.36	1.00	[41]
<i>Staphylococcus pseudintermedius</i>	407	0.24	3.97	[44]
<i>Yersinia pestis</i>	530	0.03	3.00	[41]

NHA, non-host-associated
 HA, host-associated
 -, information unavailable

catalyzes the two-step biosynthesis of anthranilate and has experienced trans-kingdom gene fusions [46]; *cstA* is involved in peptide utilization during carbon starvation [47]; *betT* encodes a choline-glycine betaine transporter, which could help to overcome osmotic stress by the accumulation of compatible solutes [48, 49]; and *comEC* encodes a transformation protein and has been shown to be absolutely required for DNA uptake and recognition [50].

Habitat-associated accessory OGs underlying putative differential adaptation

M. luteus is widely distributed in various environments and a few strains were isolated as human pathogens or from the mammalian skin or other tissues. In order to investigate whether the genomic differences between clades are relevant to the habitat, the isolation resources of all strains were mapped to the phylogenomic tree (Fig. 3 and Figure S6). Strains whose isolation sources were known could be classified into two groups: host-

associated (HA), including 18 strains isolated from mammal hosts, and non-host-associated (NHA), including 47 strains isolated from other sources or free living. HA strains were not clustered on the tree but rather widely distributed in all clades, indicating the inter-clade genetic difference was irrelevant to the habitat transition. It was noted that HA strains tended to locate on the lately diverged branches, and that strains on the early diverged branches as well as the closest characterized relatives of *M. luteus* (*M. endophyticus* and *M. flavus*) were all isolated from NHA habitats. These results suggested that the last common ancestor of *M. luteus* was more likely to be genuine dwellers in NHA niches, and a few descendants transited to a mammal-associated lifestyle recently.

Lifestyle changes are drastic events followed by gene gain and loss. We sought to study whether strains from different habitats were associated with different sets of genes. A pan-genome-wide association study (pan-GWAS) revealed 101 accessory genes that were present in at least half of the strains from one habitat but no more than half from the other. Among the 13 genes whose associations were statistically significant (Fig. 5 and Table S2, Benjamini-Hochberg *P*-value < 0.1 and Empirical *P*-value < 0.05), eleven were enriched in NHA strains and two were enriched in HA strains. It was noted that the two HA genes (OG_2452 and OG_2453) and one of the NHA genes (OG_1888) were located at the same genomic region (Fig. 6a). Genes of OG_2452 and OG_2453 encoded a sortase and an excalibur

calcium-binding domain-containing protein, respectively. Sortases are enzymes responsible for covalent anchoring of specific proteins to the peptidoglycan of the cell wall of Gram-positive bacteria, performing critical biological functions that are required for the colonization and invasion of host tissues [51, 52]. Sortases have also been considered as important virulence factors, as they play key roles in the infection process [53]. The NHA gene OG_1888 encoded a protein that was a member of COG0739 of membrane proteins related to metallo-endopeptidases. However, the function of OG_1888 in the adaption to the NHA habitats is unknown. Considering the conserved flanking genes and the presence of a transposase gene inside the region, it is possible that the integration of this region was a result of horizontal gene transfer (HGT). Actually, the NHA gene OG_1888 was located within a putative GI, further supporting the hypothesis. Additionally, the OG_2452-OG_2453 gene cluster existed in HA strains located on different clades, indicating that different HA lineages obtained this gene cluster separately and that there may be HGT between different HA lineages.

Most NHA genes were clustered within two regions. One such region contained six NHA genes (OG_1971 to OG_1976) (Fig. 6b). In this region, three genes, *fadA*, *fadB*, and *fadE*, responsible for the degradation and recycling of fatty acids via β -oxidation, were identified. Fatty acids are essential components of membranes and are important sources of metabolic energy in all organisms [54]. The other region contained genes encoding a

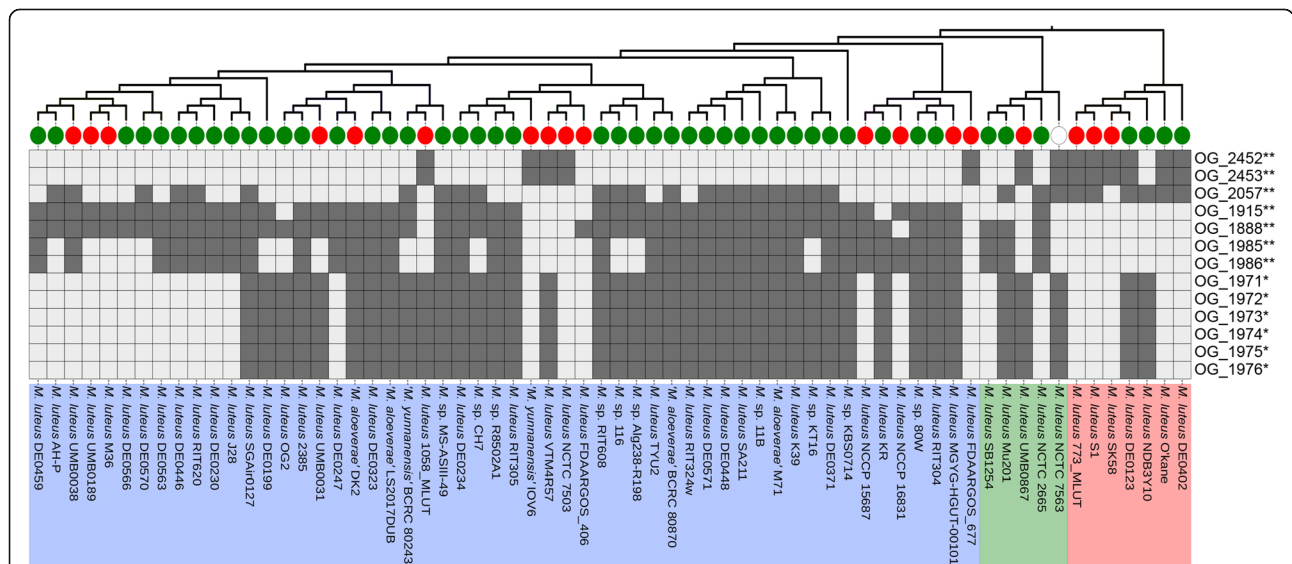
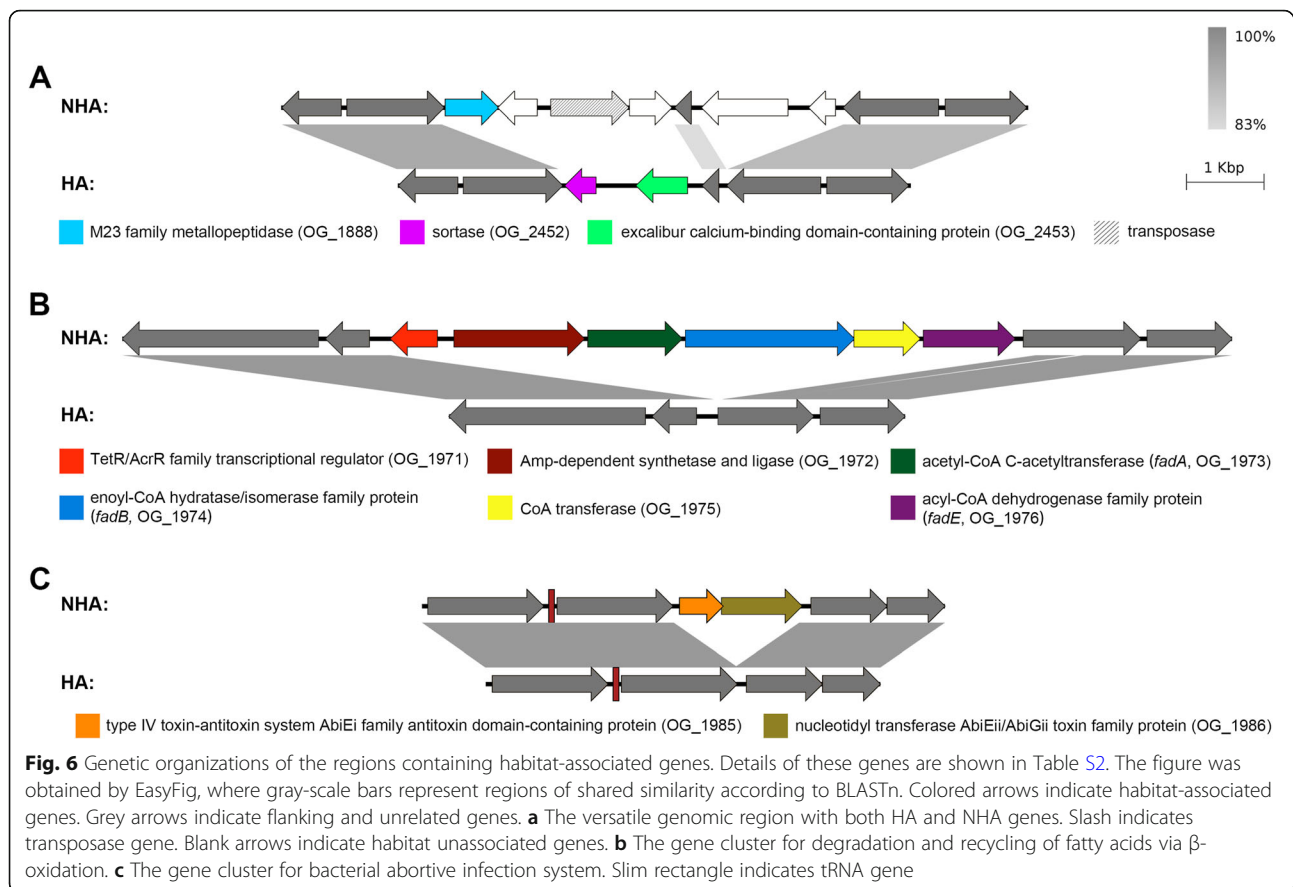


Fig. 5 Distribution of the habitat-associated OGs. Details of these genes are shown in Table S2. Heatmaps represent the distribution of 13 accessory OGs that significantly associated with HA and NHA habitats. Dark and light gray boxes indicate presence and absence, respectively. The tree topology is derived from Fig. 3, and the typical clades are marked with different colors. Strains from different habitats are represented by circles in different colors (red, HA; green, NHA; blank, uncertain). Asterisks indicate the naive *P*-value for the null hypothesis that the presence/absence of this OG is unrelated to the habitat (*, *P* < 0.05; **, *P* < 0.01)



type IV toxin-antitoxin ‘innate immunity’ bacterial abortive infection (Abi) system, consisted of a bicistronic operon encoding an AbiEi antitoxin and an AbiEii toxin (OG_1985 and OG_1986, Fig. 6c). The Abi system, abbreviated for the phage abortive infection system, provides a post-infection resistance mechanism that could block phage multiplication and result in the death of the infected bacterial cell upon phage infection [55–57]. Beside these two genomic regions, there were also two genes (OG_1915 and OG_2057) showing a significant association with NHA habitat. One of these (OG_2057) encoded a glyoxalase (COG0346/PF18029), which has been reported to play a role in the detoxification of methylglyoxal in bacterial cell metabolism when suffering environmental stresses [58]. Thus, these NHA genes probably enhanced the adaptation to complex environments.

Discussion

The ability of *M. luteus* to survive and reproduce in a wide range of ecological niches and to cause disease not only is practically significant, but also provides an ideal opportunity for inferring the population genetics and evolution of environmentally transmitted pathogens. In

fact, *M. luteus* was once considered to be a non-pathogenic saprophyte or pure contaminant from skin and mucous membranes, but it is now proved to cause severe infections in immunocompromised populations [3–5]. In this study, a genomic investigation on VF and ARG profiles of *M. luteus* was performed. It was noted that about half the ARGs and VFs were related to GIs, suggesting that HGT might be an important driving force for antibiotic resistance and pathogenicity acquisition in *M. luteus*. In addition, most of the previously reported cases due to *M. luteus* could be successfully treated with a combination of vancomycin and rifampin [59], but there was also a report that a treatment regimen consisting of vancomycin, gentamicin, and rifampicin for 4 weeks was not successful in the case of native aortic valve endocarditis secondary to *M. luteus* [60], suggesting the antibiotic resistance mechanism of *M. luteus* is still very complicated. Consistently, our analyses showed that all of the total 66 genomes contained the *rbpA* gene, the product of which has been reported to confer basal levels of rifampicin resistance on *Streptomyces coelicolor* [61]. Therefore, the VF and ARG profiles may provide guidance for the future treatment of *M. luteus* infections. However, these results may be biased

due to the incomplete genome sequences for most strains, though only high-quality genome sequences (completeness > 95% and contamination < 5%) were included in our dataset.

The present study also constitutes a comprehensive comparative genomic characterization of *M. luteus*. We show that *M. luteus* is characterized by high genomic diversity with large open pan-genome and high number of accessory and unique genes. This pan-genome diversity is considered an indicator of high HGT rates, especially for strains living in multiple niches [62, 63], suggesting that *M. luteus* had a high capacity to absorb and integrate external genetic elements from environments. This is further supported by the detection of extensive gene gain/loss events that have occurred along the evolutionary history. All the above results showed a considerable intraspecies heterogeneity of *M. luteus*, which has been preliminarily proposed based on macro-restriction analysis using pulsed-field gel electrophoresis [64]. This high genome variability may contribute to the functional diversity of *M. luteus* strains thriving in various habitats.

Phylogenomic analysis indicated that *M. luteus* has diverged into three well-differentiated clades, while population structure analysis revealed four admixed ancestral subpopulations, with three of them corresponding to the three clades, respectively, and the fourth (represented by strain NCTC 7563) corresponding to an unknown ancestor. Besides, since isolation source information for strain NCTC 7563 was missing in the current dataset, other information and more sampling are needed to clarify what the predicted POP4 really is. In summary, while the population structure analysis supported the inter-clade differences, it suggested the existence of an unknown ancestor that contributed to the diversity of the species or the existence of the fourth, yet unsampled, clade. Furthermore, consistent with the high proportion of admixture inferred from population assignment and also the numerous insertion sequences and transposases detected in the genomes, our study provides compelling evidence for high-level homologous recombination within *M. luteus*. This will probably allow frequent gene exchange between *M. luteus* strains and with other organisms. Recombination has been proved as an important driver of the evolution of most prokaryotes, and acquisition of novel alleles of existing genes will also accelerate ecological adaptation of bacterial populations [65, 66].

It has been shown that ecological niches affect the evolution of bacteria [67, 68]. Since the transition between HA and NHA lifestyles is a dramatic ecological change, adaptive evolutionary processes will occur [69]. Based on the distribution of HA and NHA *M. luteus* strains on the phylogenetic tree, we proposed that the last common ancestor of *M. luteus* was more likely to be

genuine dwellers in NHA niches. This proposition is different from an early speculation that *M. luteus* was primarily adapted to (mammalian) skin, and its presence in water or soil might come from contamination by skin flakes [12]. By using pan-GWAS, we identified a number of genes that presumably contribute to adaptation of *M. luteus* to different lifestyles, although the clustering patterns of these genes do not exactly follow habitat. For example, the sortase probably enables HA strains better colonization on mammalian hosts [52], and the Abi system may act as stress-response elements that helps NHA strains survive unfavorable growth conditions [57, 70]. Nevertheless, much of the inferred habitat-associated genes remain uncharted and/or await experimental verification. These genes suggested that the nascent ecological differentiation has already been initiated and pathogenic strains and lineages in this species are merging. However, it is unclear whether the differentiation has been completed. It has been proposed that microbial speciation is usually driven by natural selection for adaptation to distinct ecological niches, and the distinctness is further maintained by barriers to gene flow [71]. This process is not necessarily continuous and complete, and may be terminated at any time [14]. In fact, we did not detect any clues to the existence of recombination barriers between habitats, as there was no obvious difference in recombination parameters between datasets of HA, NHA, and all strains (Table 1). Here, using *M. luteus*, we have shown another case to depict the process of bacterial ecological differentiation, whether or not it proceeds to completion. During this process, niche-specifying (adaptive) variation has already emerged. However, the separation process seems to be suppressed probably because of the high-level recombination and the strong diffusibility between niches.

Conclusions

We performed comprehensive comparative genomic analyses of 66 *Micrococcus luteus* genomes. Our study revealed high intraspecies genome diversity and extensive gene gain/loss events that contributed to the genome diversity, and showed that extensive recombination played key roles in the evolution of the species. Our results suggested that the last common ancestor of *Micrococcus luteus* had a free-living lifestyle and a few lineages have recently developed or are developing a mammal-associated lifestyle independently. This recent ecological differentiation appeared to be a process that was not related to the early phylogenetic separation of different clades of the species. This study highlighted the complicated evolution history of an emerging nosocomial pathogen *Micrococcus luteus* and provided new insights into mechanisms that drive the diversification of a species and adaptation to various environments.

Methods

Genome data set

All genomes of the members of the genus *Micrococcus* ($n = 106$, April 2020) that derived from different projects and habitats were downloaded from the NCBI genomes FTP site (<https://www.ncbi.nlm.nih.gov/genomes>). In order to ensure the reliability of the results, we used CheckM [72] to assess isolate genome completeness and contamination. Only genomes that were at least 95% complete and had no more than 5% contamination were used. In addition, we computed whole genome ANI for each pair of genomes by pyani (<https://github.com/widdowquinn/pyani>) in ANIm mode. To avoid the statistical deviation caused by highly similar strains or even multiple sequencing of the same strain (e.g., strain NCTC 2665), we considered genome pair redundant when ANI was higher than 99.9%. In such cases, one genome with the most completeness and the largest N50 was selected, and the others were marked as “redundant” and were filtered out. Through this, a total of 76 high-quality and non-redundant *Micrococcus* genomes were retained. The information of isolation source of each strain was obtained from the NCBI BioSample and BioProject databases and from the literature. Similarity searches for the 16S rRNA gene sequences were performed by the EzBioCloud server [73].

Genome annotation and determination of OGs

To maintain consistency in gene annotations and to reduce the system errors caused by different programs or parameters, all genomes were reannotated locally and analyzed using the same strategy. De novo gene predictions for CDSs were performed with PRODIGAL [74]. Function annotation and classification of proteins were performed by sequence comparison against the eggNOG database [26], using the eggNOG-Mapper [75] with the DIAMOND method [76]. If a gene was assigned to more than one COG category, each COG category was calculated separately. Pfam domains were further identified by hmmscan in HMMER [77] against the PFAM-A database [78] with E-value $\leq 1e-05$. Insertion sequences and transposases were identified by BLASTp against the ISFinder database [79] with manual inspection of search hits (E-value $\leq 1e-05$). GIs were predicted using IslandPATH-DIMOB [80]. Potential ARGs and putative VFs encoded in *M. luteus* genomes were identified through BLASTp searches of the Comprehensive Antibiotic Resistance Database (CARD) [81] and the VFDB [28], respectively, with E-values $\leq 1e-05$ and *Ha*-value > 0.42 [82]. Heatmaps were created by the pheatmap package in R (<http://cran.r-project.org/web/packages/pheatmap/index.html>).

All proteins were clustered using GET_HOMOLOGUES [33] to identify OGs with the OrthoMCL

clustering algorithm. A relatively rigorous standard with 70% sequence identity and 70% coverage as minimum BLASTp homology cutoff was used. Cloud, shell, and (soft-) core pangenome components were derived according to GET_HOMOLOGUES. Pan-genome statistics were visualized using PanGP [83].

Phylogenomics and population structure analyses

For phylogenomic analyses, only single-copy core genes were used. Codon-based alignments for each OGs were obtained by aligning the translated protein with MAFFT [84] and back-translating with PAL2NAL [85]. Poorly aligned regions were filtered by Gblocks [86], with default parameters except for option $-t = c$. The aligned sequences were concatenated as a single data set using homemade Python scripts. The approximately-maximum-likelihood phylogenetic tree was built using the generalized time-reversible model implemented in FastTree [87] and visualized using the Interactive Tree of Life [88]. In addition, we used R function pvclust [89] to perform a hierarchical cluster analysis, based on an absence/presence (0/1) matrix of dispensable genes according to GET_HOMOLOGUES results. To further elucidate the population structure of *M. luteus*, we used Fastbaps [35] with default parameters based on the core genome alignment. We also used the model-based Bayesian method implemented in Structure 2.3.4 [36], in which the admixture model was used with a varying K from 2 to 10, and the optimal value for K was obtained by STRUCTURE HARVESTER [90]. The pangenome matrix and the rooted species tree were used as inputs for COUNT [91] to calculate posterior probabilities for gain and loss of each OG across all nodes during the evolution of *M. luteus*.

Estimation of genetic recombination

Four approaches were used to detect recombination in *M. luteus*. Based on the concatenated single-copy core genes, we firstly used SplitsTree [39] to construct a network with the NeighborNet algorithm, and also to calculate the pairwise homoplasy index [40]. Then, we ran fastGEAR [45] to detect recent and ancestral recombination events happening in each single-copy core gene with default parameters. In addition, PhiPack [40] was also used, by which potential recombination events were identified by having *P*-values lower than 0.05 in at least two of the three methods (PHI, Neighbor Similarity Score and Maximum Chi-Square) computed from 1000 permutations. Finally, we used mcorr [41] with default parameters to calculate correlation profiles and the recombination parameters, including the mean fragment size of a recombination event (f), the fraction of sample diversity derived from recombination (c), and the relative rate of recombination to mutation (γ/μ).

Identification of the habitat-associated genes

Firstly, two 0/1 matrices were created: one based on the presence/absence of candidate habitat-associated genes (present in at least half of the strains of one habitat but no more than half of the other), and the other based on the habitat information of each strain, respectively. Then, a pan-GWAS using SCOARY [92] was carried out with 1000 permutation replicates to identify genes that significantly associated with HA or NHA lifestyle. Genes associated with different lifestyles were identified by odds ratio, and were considered as significant only when they attained a naive *P*-value less than 0.05, an empirical *P*-value less than 0.05, and a Benjamini-Hochberg corrected *P*-value less than 0.1. Pairwise comparisons of specific genomic regions within representative *M. luteus* strains were visualized using EasyFig [93].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-07432-5>.

Additional file 1: Table S1. Genomic features and isolation sources of strains used in this study

Additional file 2: Figure S1. Variations in genome size and GC content between species in genus *Micrococcus*. **Figure S2.** Heatmap of the distributions of the putative VF genes in *M. luteus* genomes. **Figure S3.** Antibiotic resistance gene profiles of *M. luteus* strains. **Figure S4.** Comparison of the soft-core, shell and cloud genomes of *M. luteus* based on COG categories. **Figure S5.** Hierarchical cluster analysis based on the presence or absence of dispensable genes. **Figure S6.** Population structure of *M. luteus*. **Figure S7.** Ancestral genome content reconstruction using COUNT software

Additional file 3: Table S2. Accessory genes associated with HA or NHA habitats in *M. luteus*

Abbreviations

Abi: Abortive infection; ANI: Average nucleotide identity; ARG: Antibiotic resistance gene; CARD: Comprehensive Antibiotic Resistance Database; CDS: Coding sequence; COG: Clusters of Orthologous Groups; GI: Genomic island; HA: Host-associated; HGT: Horizontal gene transfer; IS: Insertion sequence; ML: Maximum-likelihood; NCBI: National Center for Biotechnology Information; NHA: Non-host-associated; OG: Orthologous group; pan-GWAS: Pan-genome-wide association study; PHI: Pairwise homoplasy index; Rpf: Resuscitation-promoting factor; VBNC: Viable but non-culturable; VF: Virulence factor; VFDB: Virulence Factors Database

Acknowledgements

Not applicable.

Authors' contributions

YSL and BBX contributed to conception and design of the study. YSL performed the main bioinformatic analyses, and wrote the draft manuscript. ZZS, JCR and BBX interpreted and discussed the results and revised the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This study was supported by the National Science Foundation of China (31770412), the Young Scholars Program of Shandong University (2016WLJH41), and Youth Interdisciplinary Science and Innovative Research Groups of Shandong University (2020QNQT006).

Availability of data and materials

All data analyzed during this study are available through NCBI GenBank database, and are accessible through the accession numbers listed in Table S1.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 August 2020 Accepted: 8 February 2021

Published online: 18 February 2021

References

1. Parte AC. LPSN - list of prokaryotic names with standing in nomenclature (bacterio.net), 20 years on. *Int J Syst Evol Microbiol.* 2018;68:1825–9.
2. Magee JT, Burnett IA, Hindmarch JM, Spencer RC. *Micrococcus* and *Stomatococcus* spp. from human infections. *J Hosp Infect.* 1990;16:67–73.
3. Khan A, Aung TT, Chaudhuri D. The first case of native mitral valve endocarditis due to *Micrococcus luteus* and review of the literature. *Case Rep Cardiol.* 2019;2019:5907319.
4. Ianniello NM, Andrade DC, Ivancic S, Eckardt PA, Lemos Ramirez JC. Native valve infective endocarditis due to *Micrococcus luteus* in a non-Hodgkin's lymphoma patient. *IDCases.* 2019;18:e00657.
5. von Eiff C, Kuhn N, Herrmann M, Weber S, Peters G. *Micrococcus luteus* as a cause of recurrent bacteremia. *Pediatr Infect Dis J.* 1996;15:711–3.
6. Kocur M, Kloos WE, Schleifer K-H. The genus *Micrococcus*. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. *The Prokaryotes: Volume 3: Archaea Bacteria: Firmicutes, Actinomycetes.* New York: Springer New York; 2006. p. 961–71.
7. Mukamolova GV, Turapov OA, Kazarian K, Telkov M, Kaprelyants AS, Kell DB, et al. The *rpf* gene of *Micrococcus luteus* encodes an essential secreted growth factor. *Mol Microbiol.* 2002;46:611–21.
8. Mukamolova GV, Murzin AG, Salina EG, Demina GR, Kell DB, Kaprelyants AS, et al. Muralytic activity of *Micrococcus luteus* Rpf and its relationship to physiological activity in promoting bacterial growth and resuscitation. *Mol Microbiol.* 2006;59:84–98.
9. Yu C, Liu Y, Jia Y, Su X, Lu L, Ding L, et al. Extracellular organic matter from *Micrococcus luteus* containing resuscitation-promoting factor in sequencing batch reactor for effective nutrient and phenol removal. *Sci Total Environ.* 2020;727:138627.
10. Su XM, Liu YD, Hashmi MZ, Ding LX, Shen CF. Culture-dependent and culture-independent characterization of potentially functional biphenyl-degrading bacterial community in response to extracellular organic matter from *Micrococcus luteus*. *Microb Biotechnol.* 2015;8:569–78.
11. Jampasri K, Pokethititook P, Kruatrachue M, Ounjai P, Kumsopa A. Phytoremediation of fuel oil and lead co-contaminated soil by *Chromolaena odorata* in association with *Micrococcus luteus*. *Int J Phytoremediation.* 2016; 18:994–1001.
12. Young M, Artsatbanov V, Beller HR, Chandra G, Chater KF, Dover LG, et al. Genome sequence of the Fleming strain of *Micrococcus luteus*, a simple free-living actinobacterium. *J Bacteriol.* 2010;192:841–60.
13. Cohan FM. Bacterial species and speciation. *Syst Biol.* 2001;50:513–24.
14. Shapiro BJ, Polz MF. Microbial speciation. *Cold Spring Harb Perspect Biol.* 2015;7:a018143.
15. Pal C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet.* 2005;37:1372–5.
16. Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet.* 2016;17:379–91.
17. Lin H, Yu M, Wang X, Zhang XH. Comparative genomic analysis reveals the evolution and environmental adaptation strategies of vibrios. *BMC Genomics.* 2018;19:135.
18. Li Y, Pinto-Tomás AA, Rong X, Cheng K, Liu M, Huang Y. Population genomics insights into adaptive evolution and ecological differentiation in streptomycetes. *Appl Environ Microbiol.* 2019;85:e02555–18.
19. Tian X, Zhang Z, Yang T, Chen M, Li J, Chen F, et al. Comparative genomics analysis of *Streptomyces* species reveals their adaptation to the marine

- environment and their diversity at the genomic level. *Front Microbiol.* 2016; 7:998.
20. Kutnutia SK, Drautz-Moses DI, Uchida A, Purbojati RW, Wong A, Kushwaha KK, et al. Complete genome sequence of *Micrococcus luteus* strain SGAir0127, isolated from indoor air samples from Singapore. *Microbiol Resour Announc.* 2019;8:e00646–19.
 21. Lee S, An YW, Choi CH, Yun MR, Kim S, Cheong H, et al. Complete genome sequences of *Micrococcus luteus* strains NCCP 15687 and NCCP 16831, isolated in South Korea. *Microbiol Resour Announc.* 2020;9:e01558–19.
 22. Lafi FF, Ramirez-Prado JS, Alam I, Bajic VB, Hirt H, Saad MM. Draft genome sequence of plant growth-promoting *Micrococcus luteus* strain K39 isolated from *Cyperus conglomeratus* in Saudi Arabia. *Genome Announc.* 2017;5:e01520–16.
 23. Chen HH, Zhao GZ, Park DJ, Zhang YQ, Xu LH, Lee JC, et al. *Micrococcus endophyticus* sp. nov., isolated from surface-sterilized *Aquilaria sinensis* roots. *Int J Syst Evol Microbiol.* 2009;59:1070–5.
 24. Rieser G, Scherer S, Wenning M. *Micrococcus cohnii* sp. nov., isolated from the air in a medical practice. *Int J Syst Evol Microbiol.* 2013;63:80–5.
 25. Liu XY, Wang BJ, Jiang CY, Liu SJ. *Micrococcus flavus* sp. nov., isolated from activated sludge in a bioreactor. *Int J Syst Evol Microbiol.* 2007;57:66–9.
 26. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47:D309–14.
 27. Vandecraen J, Chandler M, Aertsen A, Van Houdt RJCrim. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit Rev Microbiol.* 2017;43:709–30.
 28. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 2018;47:D687–92.
 29. Zhou C, Bhinderwala F, Lehman MK, Thomas VC, Chaudhari SS, Yamada KJ, et al. Urease is an essential component of the acid response network of *Staphylococcus aureus* and is required for a persistent murine kidney infection. *PLoS Pathog.* 2019;15:e1007538.
 30. Rutherford JC. The emerging role of urease as a general microbial virulence factor. *PLoS Pathog.* 2014;10:e1004062.
 31. Skaar EP, Tobiason DM, Quick J, Judd RC, Weissbach H, Etienne F, et al. The outer membrane localization of the *Neisseria gonorrhoeae* MsrA/B is involved in survival against reactive oxygen species. *Proc Natl Acad Sci U S A.* 2002;99:10108–13.
 32. Olry A, Boschi-Muller S, Marraud M, Sanglier-Cianferani S, Van Dorsselaar A, Branlant G. Characterization of the methionine sulfoxide reductase activities of PILB, a probable virulence factor from *Neisseria meningitidis*. *J Biol Chem.* 2002;277:12016–22.
 33. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol.* 2013;79:7696–701.
 34. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol.* 2005;3:711–21.
 35. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* 2019;47:5539–49.
 36. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* 2003;164:1567–87.
 37. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. *Science.* 2007;315:476–80.
 38. Dixit PD, Pang TY, Maslov S. Recombination-driven genome evolution and stability of bacterial species. *Genetics.* 2017;207:281–95.
 39. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 2006;23:254–67.
 40. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. *Genetics.* 2006;172:2665–81.
 41. Lin M, Kussell E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods.* 2019;16:199–204.
 42. Park CJ, Andam CP. Distinct but intertwined evolutionary histories of multiple *Salmonella enterica* subspecies. *mSystems.* 2020;5:e00515–9.
 43. Lee IPA, Andam CP. Pan-genome diversification and recombination in *Cronobacter sakazakii*, an opportunistic pathogen in neonates, and insights to its xerotolerant lifestyle. *BMC Microbiol.* 2019;19:306.
 44. Smith JT, Amador S, McGonagle CJ, Needle D, Gibson R, Andam CP. Population genomics of *Staphylococcus pseudintermedius* in companion animals in the United States. *Commun Biol.* 2020;3:282.
 45. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol Biol Evol.* 2017;34:1167–82.
 46. Yanai I, Wolf YI, Koonin EV. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol.* 2002;3:research0024.
 47. Schultz JE, Matin A. Molecular and functional characterization of a carbon starvation gene of *Escherichia coli*. *J Mol Biol.* 1991;218:129–40.
 48. Boscari A, Mandon K, Dupont L, Poggi MC, Le Rudulier D. BetS is a major glycine betaine/proline betaine transporter required for early osmotic adjustment in *Sinorhizobium meliloti*. *J Bacteriol.* 2002;184:2654–63.
 49. Lamark T, Kaasen I, Eshoo MW, Falkenberg P, McDougall J, Strom AR. DNA sequence and analysis of the *bet* genes encoding the osmoregulatory choline-glycine betaine pathway of *Escherichia coli*. *Mol Microbiol.* 1991;5:1049–64.
 50. Pimentel ZT, Zhang Y. Evolution of the natural transformation protein, ComEC, in bacteria. *Front Microbiol.* 2018;9:2980.
 51. Cascioferro S, Totsika M, Schillaci D. Sortase a: an ideal target for anti-virulence drug development. *Microb Pathog.* 2014;77:105–12.
 52. Cossart P, Jonquières R. Sortase, a universal target for therapeutic agents against gram-positive bacteria? *Proc Natl Acad Sci U S A.* 2000;97:5013–5.
 53. Sprig T, Weiner EM, Clubb RT. Sortase enzymes in gram-positive bacteria. *Mol Microbiol.* 2011;82:1044–59.
 54. Fujita Y, Matsuoka H, Hirooka K. Regulation of fatty acid metabolism in bacteria. *Mol Microbiol.* 2007;66:829–39.
 55. Hampton HG, Jackson SA, Fagerlund RD, Vogel AIM, Dy RL, Blower TR, et al. AbiEi binds cooperatively to the type IV *abiE* toxin-antitoxin operator via a positively-charged surface and causes DNA bending and negative autoregulation. *J Mol Biol.* 2018;430:1141–56.
 56. Dy RL, Przybicki R, Semeijn K, Salmond GP, Fineran PC. A widespread bacteriophage abortive infection system functions through a type IV toxin-antitoxin mechanism. *Nucleic Acids Res.* 2014;42:4590–605.
 57. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. *Nat Rev Microbiol.* 2010;8:317–27.
 58. MacLean MJ, Ness LS, Ferguson GP, Booth IR. The role of glyoxalase I in the detoxification of methylglyoxal and in the activation of the KefB K⁺ efflux system in *Escherichia coli*. *Mol Microbiol.* 1998;27:563–71.
 59. Rodriguez-Nava G, Mohamed A, Yanez-Bello MA, Trelles-Garcia DP. Advances in medicine and positive natural selection: prosthetic valve endocarditis due to biofilm producer *Micrococcus luteus*. *IDCases.* 2020;20:e00743.
 60. Miliadous G, Elisaf M. Native valve endocarditis due to *Micrococcus luteus*: a case report and review of the literature. *J Med Case Rep.* 2011;5:251.
 61. Newell KV, Thomas DP, Brekasis D, Paget MS. The RNA polymerase-binding protein RbpA confers basal levels of rifampicin resistance on *Streptomyces coelicolor*. *Mol Microbiol.* 2006;60:687–96.
 62. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* 2008;11:472–7.
 63. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev.* 2005;15:589–94.
 64. Murayama O, Matsuda M, Moore JE. Studies on the genomic heterogeneity of *Micrococcus luteus* strains by macro-restriction analysis using pulsed-field gel electrophoresis. *J Basic Microbiol.* 2003;43:337–40.
 65. Sun Y, Luo H. Homologous recombination in core genomes facilitates marine bacterial adaptation. *Appl Environ Microbiol.* 2018;84:e02545–17.
 66. González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. Impact of homologous recombination on the evolution of prokaryotic core genomes. *mBio.* 2019;10:e02494–18.
 67. Shapiro BJ, Leducq JB, Mallet J. What is speciation? *PLoS Genet.* 2016;12:e1005860.
 68. Vos M. A species concept for bacteria based on adaptive divergence. *Trends Microbiol.* 2011;19:1–7.
 69. Toft C, Andersson SG. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet.* 2010;11:465–75.
 70. Chopin MC, Chopin A, Bidnenko E. Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol.* 2005;8:473–9.
 71. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, et al. Population genomics of early events in the ecological differentiation of bacteria. *Science.* 2012;336:48–51.

72. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 2015;25:1043–55.
73. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol.* 2017;67:1613–7.
74. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11:119.
75. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol.* 2017;34:2115–22.
76. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12:59–60.
77. Eddy SR. Profile hidden Markov models. *Bioinformatics.* 1998;14:755–63.
78. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. *Nucleic Acids Res.* 2012;40:D290–301.
79. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference Centre for bacterial insertion sequences. *Nucleic Acids Res.* 2006;34:D32–6.
80. Bertelli C, Brinkman FSL. Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics.* 2018;34:2161–7.
81. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48:D517–25.
82. Li J, Tai C, Deng Z, Zhong W, He Y, Ou HY. VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Brief Bioinform.* 2018;19:566–74.
83. Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics.* 2014;30:1297–9.
84. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80.
85. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 2006;34:W609–12.
86. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 2000;17:540–52.
87. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One.* 2010;5:e9490.
88. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 2016;44:W242–5.
89. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* 2006;22:1540–2.
90. Earl D, Vonholdt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour.* 2011;4:359–61.
91. Csűrös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics.* 2010;26:1910–2.
92. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 2016;17:238.
93. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics.* 2011;27:1009–10.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

