Global diversity of enterococci and description of 18 previously unknown species

Schwartzman, Julia A.; Lebreton, Francois; Salamzade, Rauf; Shea, Terrance; Martin, Melissa J.; Schaufler, Katharina; Urhan, Aysun; Abeel, Thomas; Camargo, Ilana L.B.C.; More Authors

DOI
10.1073/pnas.2310852121

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the National Academy of Sciences of the United States of America

Citation (APA)
Schwartzman, J. A., Lebreton, F., Salamzade, R., Shea, T., Martin, M. J., Schaufler, K., Urhan, A., Abeel, T., Camargo, I. L. B. C., & More Authors (2024). Global diversity of enterococci and description of 18 previously unknown species. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(10), Article e2310852121. https://doi.org/10.1073/pnas.2310852121

PNAS RESEARCH ARTICLE | MICROBIOLOGY OPEN ACCESS

Global diversity of enterococci and description of 18 previously unknown species

Julia A. Schwartzman[a,b,c,1], Francois Lebreton[a,b,d,1], Rauf Salamzade[e,f], Terrance Shea[e], Melissa J. Martin[a,b,d], Katharina Schaufler[a,b,g,h], Aysun Urhan[e,i], Thomas Abeel[e,i], Ilana L. B. C. Camargo[j], Bruna F. Sgardioli[j], Janira Prichula[a,b,k], Ana Paula Guedes Frazzon[a,b,l], Gonzalo Giribet[m], Daria Van Tyne[a,b,n], Gregg Treinish[o], Charles J. Innis[p], Jaap A. Wagenaar[q], Ryan M. Whipple[a,b], Abigail L. Manson[e,1], Ashlee M. Earl[e,1,2], and Michael S. Gilmore[a,b,2]

Edited by Roy Curtiss III, University of Florida, Gainesville, FL; received July 10, 2023; accepted December 6, 2023

Enterococci are gut microbes of most land animals. Likely appearing first in the guts of arthropods as they moved onto land, they diversified over hundreds of millions of years adapting to evolving hosts and host diets. Over 60 enterococcal species are now known. Two species, *Enterococcus faecalis* and *Enterococcus faecium,* are common constituents of the human microbiome. They are also now leading causes of multidrug-resistant hospital-associated infection. The basis for host association of enterococcal species is unknown. To begin identifying traits that drive host association, we collected 886 enterococcal strains from widely diverse hosts, ecologies, and geographies. This identified 18 previously undescribed species expanding genus diversity by >25%. These species harbor diverse genes including toxins and systems for detoxification and resource acquisition. *Enterococcus faecalis* and *E. faecium* were isolated from diverse hosts highlighting their generalist properties. Most other species showed a more restricted distribution indicative of specialized host association. The expanded species diversity permitted the *Enterococcus* genus phylogeny to be viewed with unprecedented resolution, allowing features to be identified that distinguish its four deeply rooted clades, and the entry of genes associated with range expansion such as B-vitamin biosynthesis and flagellar motility to be mapped to the phylogeny. This work provides an unprecedentedly broad and deep view of the genus *Enterococcus*, including insights into its evolution, potential new threats to human health, and where substantial additional enterococcal diversity is likely to be found.

Enterococcus | genomics | global diversity | antibiotic resistance | host microbe interaction

Significance

Enterococci are among the most widely distributed microbes in animal gut consortia. Over 60 species have been identified, including *Enterococcus faecalis* and *Enterococcus faecium* commonly found in the human gut. Importantly, both emerged in the antibiotic era as leading causes of multidrug-resistant infection. Microbial traits that determine membership in microbiomes of various hosts are largely unknown and enterococci represent a unique opportunity to determine core underlying principles of that association. This study examined 886 enterococcal specimens from a wide range of hosts in diverse geographies and ecologies. Generalist to specialist enterococcal species were found including 18 previously undescribed species. This study identified new features associated with species radiations and provides evidence that tremendous genetic diversity in *Enterococcus* remains to be discovered.

Enterococci are unusually rugged and environmentally persistent microbes (1). This unusual hardiness appears to have aided their transmission among arthropods and then tetrapods as animals began to colonize land and now contributes to the spread of antibiotic-resistant enterococci in hospitals (2). Likely because of their emergence in the early days of terrestrialization, enterococci are among the most widely distributed members of gut microbiomes in land animals—from invertebrates to humans (3). Their occurrence in animals widely varying in gut physiologies, diets, and social habits provides a unique opportunity to explore how diverse host backgrounds gut drive microbiome membership.

Pioneering surveys in the 1960s and 1970s by Mundt and colleagues provided early evidence for the widespread occurrence of enterococci in diverse hosts including mammals and birds (4), insects (5), and animal-inhabited environments (6, 7). However, at the time the genus *Enterococcus* had yet to be recognized as distinct from *Streptococcus* (8) and was resolved into species at low resolution by a small number of metabolic tests (9). Although evidence of diversity among the enterococci was found, the inability to precisely assess species and strain differences limited the ability to associate well-defined *Enterococcus* species with particular hosts. Genomics now provides a high-resolution tool capable of detecting differing traits between species of microbes from varying hosts and for quantifying the extent of their divergence.

The goal of the current study was to sample the Earth broadly for enterococci from diverse hosts, geographies, and environments, to gain a first approximation of the diversity of species on the planet, and to compare the content and degree of divergence of their genomes toward the broader goal of understanding the mechanisms that drive association with particular hosts. To achieve these goals, 430 enterococci from 381 unprocessed animal samples, as well as 456 enterococci isolated by contributors from diverse sources, were collected and taxonomically identified at the DNA sequence level, creating a collection of 886 isolates. The entire genomes of strains exhibiting sequence diversity suggestive of distant relationship to any known species, were then sequenced in their entirety. This identified 18 previously undescribed species of *Enterococcus* and 1 new species of the ancestrally related genus

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.A.S., F.L., A.L.M., and A.M.E. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: aearl@broadinstitute.org or michael_gilmore@meei.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2310852121/-/DCSupplemental.

Published February 28, 2024.

Vagococcus. Genome sequence analysis also showed that substantial enterococcal diversity remains to be discovered, most prominently in arthropod hosts and insectivores. Further, genetic novelty was found not only in novel species but also circulating in well-known *Enterococcus* species, including a divergent BoNT-type toxin (10) and a new family of pore-forming toxins (11), highlighting the importance of broader knowledge of the enterococcal gene pool.

Results

Broad Survey Samples *Enterococcus* Host Diversity. To understand the breadth of enterococcal diversity, we examined little-sampled (non-clinical, non-human) environments, including those minimally impacted by human habitation or pollution. To maximize global coverage, we assembled the Enterococcal Diversity Consortium (EDC), an international group of scientists and adventurers, who contributed 456 colony-purified presumptive enterococci and 579 whole specimens, typically insects and scat, in addition to commercially procured samples. Enterococci were enriched from 381 of these 579 whole specimens (Dataset S1). Although sampling began prior to the Nagoya Protocol entering into force October 2014 with the objective of fair and equitable sharing of benefits arising from the utilization of genetic resources, thereby contributing to the conservation and sustainable use of biodiversity, all strains showing novelty as described below were registered with the country of origin irrespective of isolation date. Specimens derived from a wide range of hosts and host diets (e.g., carnivores versus herbivores), geographies, and environments (e.g., captive versus wild). The diversity of sources spanned penguins migrating through sub-Antarctic waters (12, 13), duiker and elephants from Uganda; insects, bivalves, sea turtles, and wild turkeys from Brazil to the United States; kestrel and vultures from Mongolia; wallaby, swans, and wombats from Australia; and zoo animals and wild birds from Europe. Two selective media, CHROMagar Orientation agar (14) and bile-esculin azide agar (9), were used to culture presumptive enterococci to minimize potential selection bias against natural enterococcal isolates with unknown properties. To recover enterococci potentially present in low abundance, we performed isolations both directly from samples and from enrichment cultures. Some specimens yielded presumptive enterococcal colonies with varying morphologies, and in those cases, each morphotype was separately analyzed (Dataset S1). At least one presumptive enterococcal isolate was culturable from 55% of samples (318 of 579), including extracts from dead insects, wild animal feces, tissue swabs, or samples of water and soil likely contaminated with animal fecal matter. These 318 presumptive *Enterococcus*-positive samples yielded 430 morphologically distinct colony types that were further analyzed. Together with the 456 putatively enterococcal isolates contributed as pure cultures by EDC members, this resulted in an analytical set of 886 presumptive enterococci (Dataset S1, Tab 1) derived from 774 different sample sources.

Preliminary Screen for Species Diversity. Species-level diversity within the genus *Enterococcus* is not well resolved by nucleotide polymorphisms in the 16S rRNA gene (15). Thus, as an initial measure of taxonomic diversity, we developed a high-resolution PCR amplification and amplicon sequencing protocol. An internal highly polymorphic 97bp fragment of an RNA methylase gene (*SI Appendix*, Fig. S1*A*), designated EF1984 in the *Enterococcus faecalis* V583 genome and previously found to be core to the *Enterococcus* genus (2), flanked by short conserved sequence stretches that could be used for amplification, was selected for initial screening for strain diversity. Sequence variation within this 97bp diversity locus (DL) proved better able to discriminate species than variability within the entire 16S rRNA gene (horizontal lines, *SI Appendix*, Fig. S1*B*). Further, sequence polymorphism within the DL recapitulated well the phylogenetic relationships between the 47 known enterococcal species based on genome-wide average nucleotide identity (ANI) (*SI Appendix*, Fig. S1*C*).

DL Variation Presumptively Identified Novel Species. Colonies of all 886 presumptive enterococcal isolates were subjected to DL amplification and amplicon sequencing. Of those, 34 isolates yielded no product. Amplification and sequencing of the full-length 16S rRNA gene of those 34 showed them to be *Carnobacterium* (9/34), *Lactobacillus* (10/34), or *Vagococcus* (14/34), all closely related to *Enterococcus*, likely accounting for their growth on selective media.

DL positive enterococci (852 of 886 isolates) derived from 41 taxonomic orders of animal hosts from 16 countries on 6 continents, representing many climatic zones (Fig. 1*A*). Largely reflecting representation in the collection, positively identified enterococcal isolates were obtained from mammals (29%), birds (28%), insects (18%), reptiles and amphibians (9%), coastal fish (4%), bivalves (2%), and gastropod mollusks (2%) (Fig. 1*B*). Samples derived from primary consumers (e.g., herbivores) as well as predators and scavengers (Fig. 1*C*). Over half of the isolates (53%) derived from wild environments with very low human activity (Fig. 1*D*).

Strains belonging to the same known species all shared DL sequence variations of 4bp or less, and DL sequence variation was able to resolve known fine-scale differences between *Enterococcus faecium* clades A and B (known to share ~94% ANI, (16, 17). Most DL sequences (96%, 824/853) matched with fewer than 4 single-nucleotide polymorphisms (SNPs) to 1 of 65 previously identified enterococcal species for which a genome had been sequenced (Fig. 2*A* and Dataset S1), to which these closely related isolates were presumptively assigned. The most frequently encountered species in our collection were: *E. faecalis* (340/853; 40%), *E. faecium* (125/853; 15% of isolates, including 66 [9%] clade A and 59 [7%] clade B), *E. mundtii* (119/853; 13%), *E. casseliflavus* (81/853;10%), and *Enterococcus hirae* (68/853; 8%). Importantly, this approach identified 27 isolates possessing 19 different DL sequences that exceeded the threshold for likely membership in a known species (>4 SNPs from any known species), identifying them as potentially novel (Fig. 2*B* and Dataset S1). These 27 isolates derived from insects (14 isolates), birds (9 isolates), and herbivorous reptiles (4 isolates) (Fig. 2*C*). In contrast, isolates of the most frequently encountered species—*E. faecalis, E. faecium*, and *E. mundtii*—derived mainly from mammals and birds, but were not exclusive to those hosts (Fig. 2*C*).

Species Boundaries and Placement within the Clade Structure of *Enterococcus*. Genomes of representative isolates that varied by more than four SNPs in the DL locus from their closest relative (n = 22) as well as 16 *Enterococcus* and 9 *Vagococcus* isolates for which few genomes occur in public databases, were sequenced to generate high-quality draft genome assemblies. A subset of 10 enterococci were also sequenced using long-read technology to generate more complete assemblies (Dataset S2). Presumptively diverse isolates that duplicated others exactly in DL sequence and derived from the same host sample type/sample location were not sequenced, as these were likely derived from the same population within the host and hence represent the same strain or genetic lineage. For each of the 47 newly sequenced *Enterococcus* and *Vagococcus* genomes,

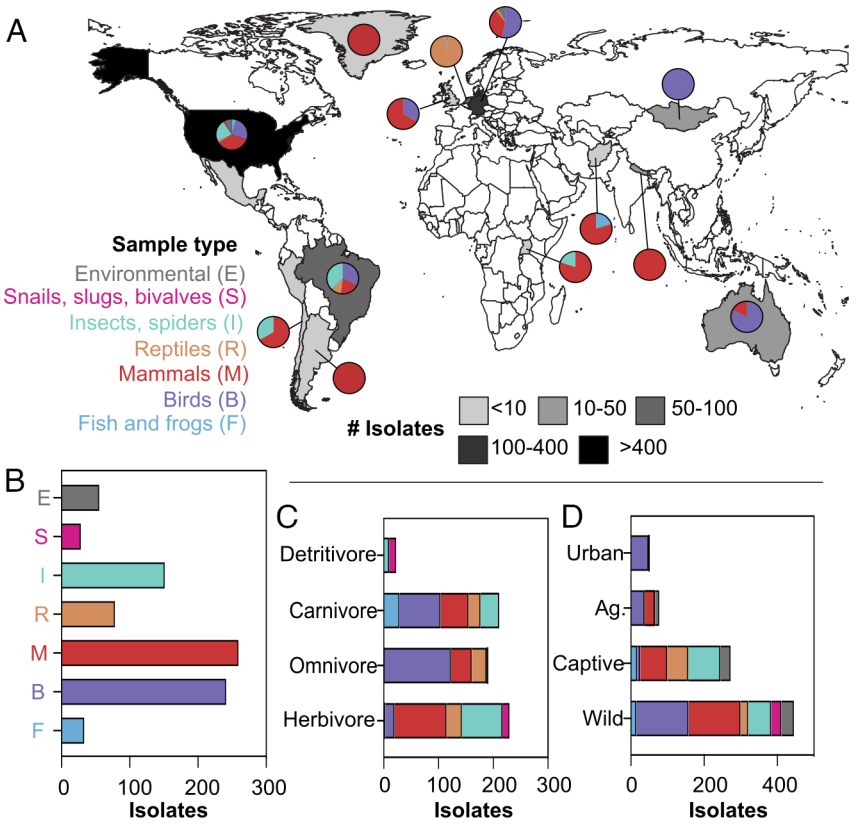

Fig. 1. Sources for 852 *Enterococcus* isolates. (*A*) Isolates derived from 16 countries, with the largest numbers collected in the United States, Germany, the Netherlands, and Brazil as reflected by gray shading. Pie chart *Insets* show the host taxonomic distribution, colored as indicated in the key. (*B–D*) Distribution of host-derived samples grouped by taxonomy, diet, and environment. Colors are the same as in panel (*A*). (*B*) Enterococcal isolates were obtained from environmental sources, as well as from 11 classes of animals representing 3 phyla: mollusks (snails, slugs, bivalves), arthropods (insects, millipedes, spiders), and vertebrates (reptiles, mammals, birds, fish, and frogs). The common names of major classes are shown with metadata in Dataset S1. (*C*) Enteric samples and feces were derived from animals with diverse diets. Where known, the diet of the animal was classified based on trophic role: detritivore, carnivore, omnivore, or herbivore. (*D*) Samples derived from locations differing in human impact. Habitats were defined as: wild, no human activity; captive, animals housed in enclosures with human contact excluding agricultural animals; agricultural animals housed in enclosures and bred by humans for agricultural purposes (Ag); urban, wild animals living in cities or towns.

and for high-quality existing assemblies of 65 known *Enterococcus* species and 4 known *Vagococcus* species, we calculated pairwise genome-wide ANI to determine species identities and explore species boundaries more accurately. The median % ANI shared by putative novel species and their nearest taxonomic neighbor was 83%, indicating that most novel species identified in this study were substantially distant and not closely related to known species. The number of DL SNPs correlated highly with genome-wide ANI (*SI Appendix*, Fig. S1*C*), validating DL sequence as an accurate predictor of novelty; 96% of isolates with >4 SNPs in the DL exhibited <95% ANI [the operationally defined species boundary (18)] with the genome of the closest identifiable species. Exceptions included two divergent *E. casseliflavus* genomes that differed by five DL SNPs from the reference but shared nominally >95% ANI (DIV0233, 95.2% ANI; 3H8_DIV0648, 95.5% ANI) with the *E. casseliflavus* reference genome. Additionally, despite differing from the reference DL by only two SNPs, three strains shared only 94.1% ANI with the reference *E. rotai* genome

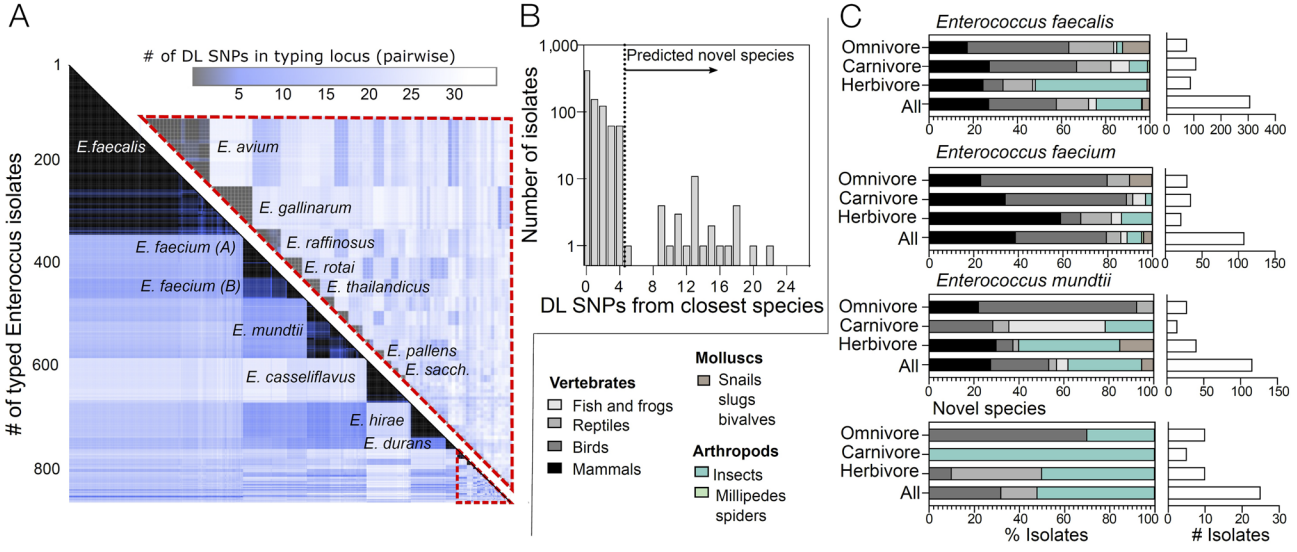

Fig. 2. Diversity of *Enterococcus* species isolated from host classes. (*A*) Heat map displaying the pairwise relatedness of the 852 putative *Enterococcus* isolates characterized by this study. Color represents the number of SNPs in the DL locus between pairs of isolates. The red dashed enclosed triangle highlights the identity of rarely encountered isolates expanded in the *Inset*. (*B*) Histogram of isolates binned by the number of SNPs in the DL locus, compared to the closest known species. The vertical dashed line is placed greater than four SNPs, the threshold chosen to prioritize genome sequencing of diverse isolates and initially predict novel species. (*C*) Host derivation of isolates, highlighting isolation sources for the three most abundant species sampled in this collection, as well as for diverse novel species.

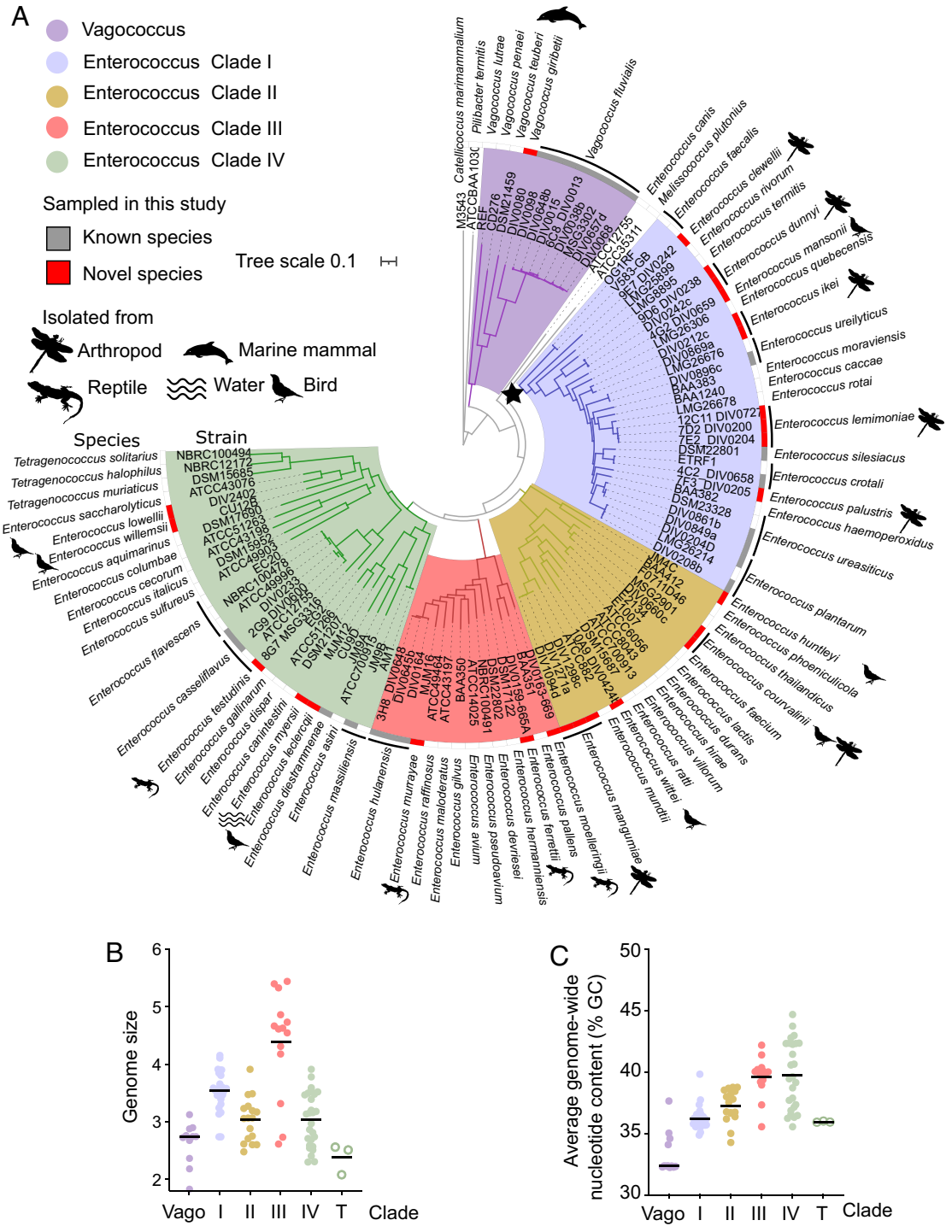

Fig. 3. Phylogeny of known and previously unknown *Enterococcus* species. (*A*) Phylogenic positioning of 5 *Vagococcus* species and 65 *Enterococcus* species, including 18 previously undescribed species of *Enterococcus* and 1 of *Vagococcus*. Black icons show the host source from which novel species isolates were derived. Strains of the same species are indicated by black lines. Clades are denoted by colored regions, Clade I blue; Clade II yellow; Clade III red; and Clade IV green. *Vagococcus* spp. are noted by purple edges. Star indicates bootstrap support ≤90%. (*B* and *C*) Distribution of genome size (*B*) and average %GC content (*C*) in our whole genome assemblies within each clade (Vago, *Vagococcus;* I–IV, *Enterococcus* Clades I–IV; T, *Tetragenococcus,* a group of divergent species clustering within *Enterococcus* Clade IV). Bars indicate the mean.

indicating that although the DL locus is conserved in this case, these also represent new species by ANI criteria. This shows that the screening criteria set were highly specific (96% of isolates with >4 SNPs in DL positive for new species) but may have been conservative. Of the *Vagococcus* genomes that were sequenced, 8/9 were confirmed to be strains of *Vagococcus fluvialis*, sharing between 97.3% and 99.9% ANI with the *V. fluvialis* reference genome (Dataset S2). One *Vagococcus* isolate was highly divergent from any known species, but taxonomically nested within the genus. This isolate, collected from infected tissue of a harbor porpoise (Dataset S2), shared only 78.5% ANI with the closest known species, *Vagococcus teuberi*, indicating that it also is a candidate novel *Vagococcus* species.

We previously observed that species within the genus *Enterococcus* clustered into four deeply branching and well-resolved clades representing early divisions, with each radiating many new species in parallel (2). Since that observation was based on a much more limited dataset, we re-evaluated enterococcal phylogenetic structure, including placement of the newly described variants and species found here, along with publicly available reference enterococcal genomes added since our prior analysis. We further included an expanded set of outgroup genomes, representing close taxonomic relatives of enterococci, including *Vagococcus, Pilibacter*, and *Catellicoccus*, including 9 *Vagococcus* genomes sequenced in this study (Dataset S2 and Fig. 3*A*).

Phylogenetic placement was based on sequence conservation in 320 single-copy genes present in all genomes. This analysis showed that the closest relative of *Enterococcus* is *Vagococcus*, with *Pilibacter* and *Catellicoccus* branching earlier within the family Enterococcaceae (Fig. 3*A*). Our sequencing of vagococci revealed a novel species sharing only 71% genome-wide ANI with *Vagococcus teuberi* (Table 1). The novel species, which we named *Vagococcus giribetii*, shares a most recent common ancestor with *V. fluvialis,* and fills in more recent diversification within the genus (Fig. 3*A*). Despite the placement of 41 additional species in the *Enterococcus* genus (Fig. 3*A*), the previously observed (2) structure of 4 deeply branching clades of the *Enterococcus* genus was supported (*Methods*). Bootstrap support for this topology was strong (≥90%), with the exception of the placement of *Melissococcus plutonius* (67%) as the most ancestral branch in clade I (Fig. 3*A*).

The long branch length of *M. plutonius* and its small genome size (2.05 Mb) reflect its unusual trajectory in becoming a highly specialized bee pathogen (19). This fundamental habitat shift within a genus of largely commensal gut microbes, likely skews its positioning in the tree and weakens its bootstrap value (67%). The higher resolution phylogeny also places *Enterococcus canis* as an outlier, branching very early in the evolution of the genus (Fig. 3*A*). All 18 previously undescribed enterococcal species in our collection nested within these four main radiations of *Enterococcus* species. Ten of these novel genomes were placed into the phylogeny at ancestral branch points relative to other species

Table 1. Candidate novel species described in this study

Species name	Strains	Clade	Most related species	ANI (%)	Host(s)	Novel/ unique genes	Examples of unique functions present in novel genomes
Enterococcus clewellii	9E7_DIV0242	I	*Enterococcus quebecensis*	<77	Captive African bullet cockroach	233	Cluster of nine genes involved in tellurium resistance, CAMP-factor toxin
Enterococcus dunnyi	9D6_DIV0238; DIV0242_7C1	I	*Enterococcus termitis*	82.0	Ground beetle Captive African bullet cockroach	49	E1-E2 ATPase
Enterococcus mansonii	4G2_DIV0659	I	*E. quebecensis*	82.3	Turkey	69	Delta-conotoxin like protein
Enterococcus ikei	DIV0869a; DIV0212c	I	*Enterococcus ureilyticus*	83.0	Captive moth Dragonfly	14	Miraculin-like protein
Enterococcus lemimoniae	12C11_DIV0727; 7D2_DIV0200; 7E2_DIV0204	I	*Enterococcus rotai*	94.1	Moth Dragonfly Dragonfly	20	Glyoxalase-like domain enzymes
Enterococcus palustris	7F3_DIV0205d	I	*Enterococcus haemoperoxidus*	84.8	Dragonfly	33	Spherulin-like cytosolic protein
Enterococcus huntleyi	JM4C	II	*Enterococcus phoeniculicola*	80.2	Chicken	83	Macrolide export-like proteins
Enterococcus courvalinii	MSG2901; DIV0660c	II	*Enterococcus thailandicus*	90.8	Cockroach Turkey	3	Restriction endonuclease, AseI-like
Enterococcus wittei	10A9_DIV0425	II	*Enterococcus mundtii*	80.6	Chicken	90	Lartarcin-like toxin, cytolysin-like toxin, Gallidermin-like protein
Enterococcus mangumiae	DIV1094d; DIV1271a; DIV1298c	II	*E. mundtii*	87.5	Butterfly Dragonfly Butterfly	38	PAAR domain protein (possible type VI secreted effector)
Enterococcus moelleringii	DIV0163-669	III	*Enterococcus pallens*	80.7	Kemp's ridley sea turtle	131	Acetoacetate decarboxylase and collangenase
Enterococcus ferrettii	DIV0159-665A	III	*E. pallens*	84.6	Kemp's ridley sea turtle	109	MAPKK protease toxin-like protein
Enterococcus murrayae	MJM16	III	*Enterococcus hulanensis*	83.7	Kemp's ridley sea turtle	80	Cluster of five genes involved in phosphonate metabolism (pnhGHIJK)
Enterococcus leclercqii	CU9D	IV	*Enterococcus diestrammenae*	79.2	Chicken	84	Alo3-like protein (antifungal), AlwI-like restriction endonuclease
Enterococcus myersii	DIV0106b-MJM12	IV	*Enterococcus dispar*	83.6	Aquarium water	50	Bacteriocin class II with double-glycine leader peptide
Enterococcus testudinis	8G7_MSG3316	IV	*Enterococcus casseliflavus*	80.6	Sea turtle	69	Putative arabinoxylan degradation, macrolide efflux, Lactococcin_972-like bacteriocin
Enterococcus willemsii	CU12B	IV	*Enterococcus saccharolyticus*	80.1	Chicken	58	Possible Beta-lactamase inhibitor gene cluster
Enterococcus lowellii	DIV2402	IV	*E. saccharolyticus*	85.4	Captive cockatoo	63	Imelysin-like secreted iron uptake system, extracellular polysaccharide biosynthesis locus
Vagococcus giribetii	DIV0080	Vago	*Vagococcus teuberi*	71.4	Harbor porpoise infection	NA	NA

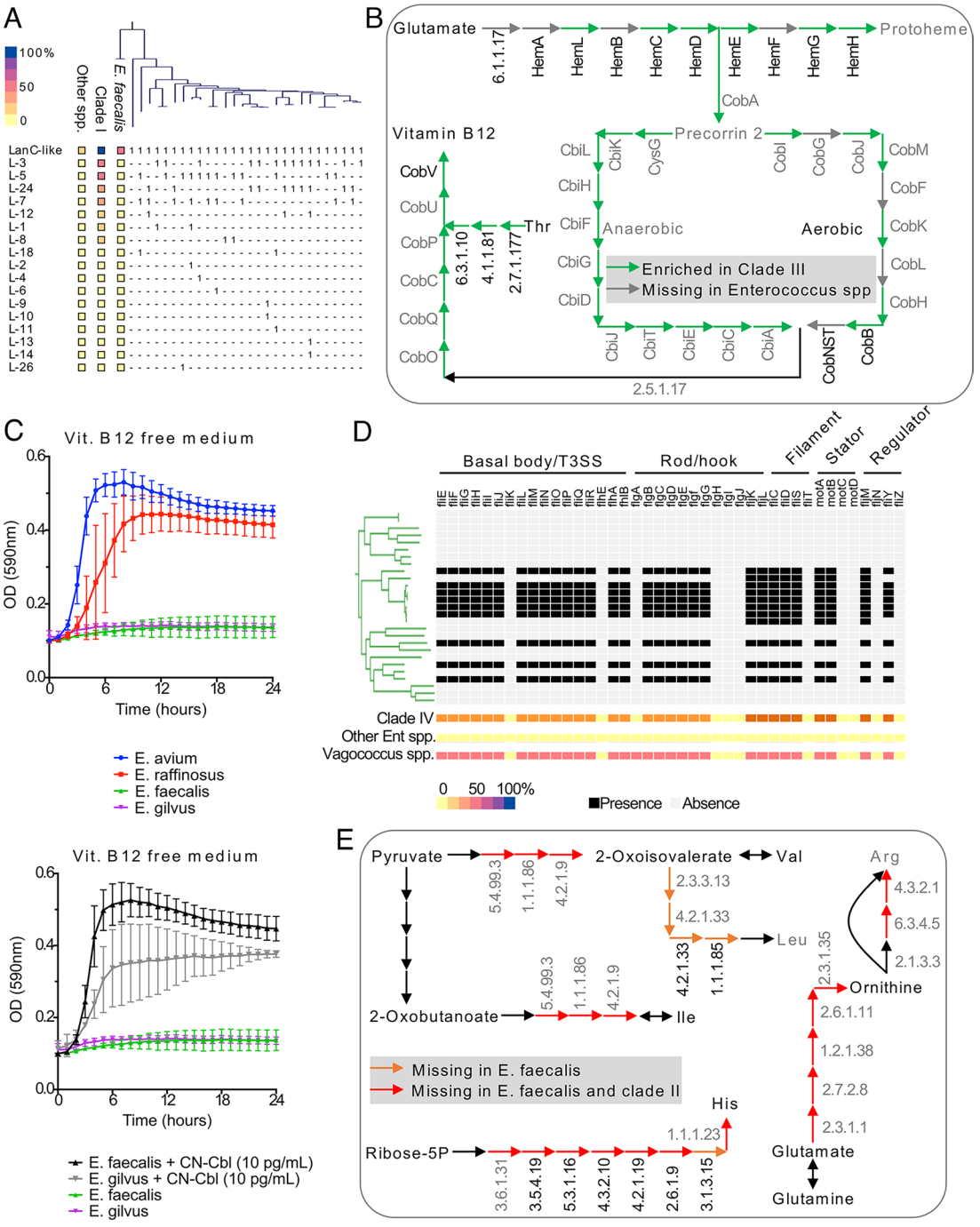

Fig. 4. Identification of clade-specific genotypic and phenotypic traits within the *Enterococcus* genus phylogeny. (*A*) Frequency of a LanC ortholog (LanC-like) as well as 17 distinct lanthipeptide synthesis clusters (Dataset S4) for Clade I species, non-Clade I *Enterococcus* (other spp.), and *E. faecalis*. The distribution of these orthologs and gene clusters is noted for individual members of Clade I as a heat map. The taxonomic relatedness of Clade I species is displayed as a pruned phylogenetic tree above the heatmap. The heatmap shows the relative abundance in *E. faecalis*, all Clade I species or all other *Enterococcus* spp. (*B*) Enrichment of complete aerobic and anaerobic cobalamin biosynthesis pathways in a subset of Clade III species. (*C*) Growth curves for a selection of isolates and species in a minimal media lacking Vitamin B12 (*Upper* chart). For species lacking the cobalamin biosynthesis pathway, growth is rescued when 10 pg/mL of cyanocobalamin (CN-Cbl) is added exogenously (*Lower* chart). (*D*) Enrichment of flagellar biosynthesis gene clusters in subset of Clade IV species. Presence (black) and absence (gray) of specific genes is indicated for each Clade IV species (displayed as a pruned phylogenetic tree) as well as a heatmap showing their prevalence in all Clade IV species, all other *Enterococcus* spp., or a set of 12 *Vagococcus* spp. (*E*) Depletion of near-complete pathways for the biosynthesis of histidine and BCAA in Clade II species together with *E. faecalis* (red arrows) or in *E. faecalis* alone (orange) compared to other *Enterococcus* spp. When available, the E.C. number of the missing enzyme is indicated.

in the topology, meaning that we expanded the known diversity in each clade (2) and revealed both recent and more evolutionarily ancient diversification within clades.

Fundamental Differences between *Enterococcus* Clades. With each of the 4 deep branching *Enterococcus* clades now represented by 12 or more species, we next sought to identify fundamental differences between clades to gain insight into the possible evolutionary drivers that created this phylogenetic structure. We previously noted that members of Clade III possessed unusually large genomes up to 5.4 Mb in size (*E. pallens*), whereas members of Clade IV possessed genomes less than half that size (*Enterococcus sulfureus*, 2.3 Mb) (2). Here, we found that Clade I genomes ranged from 2.7 Mb *E. faecalis* to the 4.2 Mb *Enterococcus termitis*, with a mean size of 3.5 Mb (Fig. 3*B*). Clade II genomes ranged in size from 2.5 Mb *Enterococcus ratti* to 3.9 Mb *E. phoeniculicola*, with a mean of 3.0 Mb. Clade III genomes, previously thought to contain only large genomes (2), were revealed to range in size from the 2.6 Mb *Enterococcus hermannienesis* to *E. pallens* (5.4 Mb), with a mean of 4.4 Mb (Fig. 3*B*). Clade IV genomes ranged in size from the previously noted 2.3 Mb *E. sulfureus* to 3.8 Mb *E. testudinis* and averaged 3.0 Mb (Fig. 3*B*). *Tetragenococcus* genomes, which grouped within Clade IV but appear to have adapted to an ecologically distinct environment from other enterococci, were slightly smaller than the average genome within Clade IV, ranging in size from 2.1 to 2.5 Mb, with a mean size of 2.4 Mb. Interestingly, except for the anomalous pathogen, *M. plutonius*, all four *Enterococcus* clades exhibited a stepwise increase in mean G+C content compared to that of ancestrally related *Vagococcus* (33%): *Enterococcus* Clade I averaged 36.2 ± 0.9%, Clade II averages 37.3 ± 1.3%, Clade III

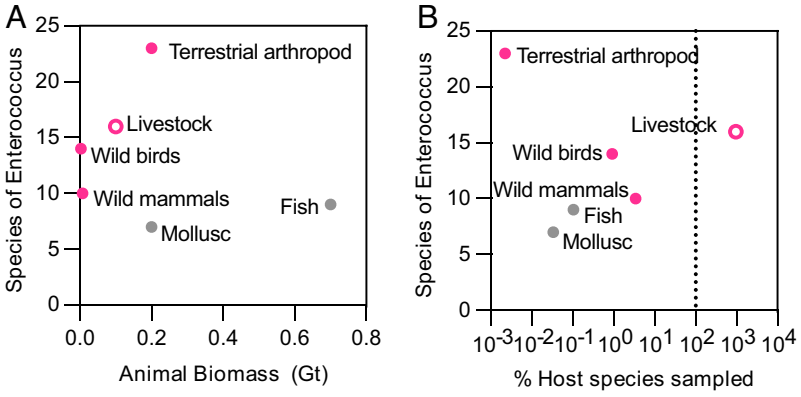

Fig. 5. Estimate of enterococcal species abundance and distribution across animal hosts. (*A*) Number of *Enterococcus* species identified in each host type, plotted as a function of total global animal biomass that host type represents. (*B*) The number of *Enterococcus* species identified as a function of the fraction of species diversity within each category sampled, showing that terrestrial livestock were comparatively oversampled relative to their diversity, whereas the diversity of terrestrial arthropod species so far has only been superficially explored. The dashed vertical line indicates 100% of diversity, with greater values representing oversampling. In both plots: pink filled circle, terrestrial; pink open circle, terrestrial-agricultural; gray circle, aquatic.

averages 39.6 ± 1.4%, and Clade IV averages 39.8 ± 2.7% GC (Fig. 3*C*).

To further characterize fundamental differences between clades, gene content was compared. The pan-genome of the *Enterococcus* genus clustered in a total of 11,086 groups of orthologous genes, excluding singletons (n = 8,017). A subset of 1,336 genes/orthogroups were shared by most (>80%) species (Dataset S3, Tab 1), while 417 genes present in single copy in all genomes of *Enterococcus* species composed the strict, single copy core (SCC) genome of the genus (Dataset S3, Tab 2). When compared to *Vagococcus*, only five members of the SCC were generally missing in species of that genus (EF1063, EF1150, EF2073, EF2149, EF2695; Dataset S3, Tab 3). Within *Enterococcus*, 67 orthogroups were enriched in Clade I and generally missing in species in Clades II to IV. Although the most commonly encountered species in this study, *E. faecalis* is clearly anomalous within Clade I as it carries genes for only 21 of the 67 orthologous groups enriched in species of that clade. Of note, all twenty-nine Clade I species, including *E. faecalis*, encoded one or more LanC-like lanthipeptide synthetase genes orthologous to *E. faecalis cylM* (EF0527), a hallmark of lanthipeptide bacteriocin biosynthetic operons that was otherwise present in only 12% of species from other clades (Fig. 4*A*). Further, 17 distinct clusters for the biosynthesis of lanthipeptides were uniquely found in Clade I with some (e.g., clusters L3, -5, -7, -24) shared by a large fraction of Clade I species (Fig. 4*A* and Dataset S4). Only five lanthipeptide synthesis clusters were found in species from other clades and, in all cases, were specific to a given isolate (Dataset S4). Besides enrichment, Clade I was depleted of 16 orthologous groups otherwise core to the other clades of *Enterococcus*. Most are of unknown function despite forming potential functional clusters/operons (e.g., E2134_0962 – E2134_0964, Dataset S3).

Clade II is characterized by the smallest core genome, and species within it generally lacked 42 orthologous genes relative to all other clades. Within those possessing a functional annotation, a large fraction of missing genes (~30 %, including operon BAA382_0711–BAA382_0714) belong to missing amino acid biosynthesis and interconversion pathways, specifically for the biosynthesis of all three branched-chain amino acids (BCAAs), but also histidine and arginine (Fig. 4*E*). Of note, the majority (>80 %) of orthologous groups depleted in Clade II were also missing in *E. faecalis* despite being core to other species in Clade I, further supporting *E. faecalis* as an outlier in this phylogenetic group and highlighting a curious convergent evolution of properties more commonly found in Clade II species. For Clades II, a total of 24 orthologous groups were significantly enriched. Most are of unknown function; however, some appear to aggregate in genomically co-located clusters (e.g., E2134_1832–E2134_1834 in *E. faecium* and most other Clade II members, Dataset S3).

Clade III was found to be significantly enriched in 82 orthologous groups, while being depleted in 8 ortholog groups relative to other enterococci. Most genes in both sets were annotated as of unknown function (Dataset S3). While not present within the arbitrary threshold of >80% of species, a large subset (64%) of Clade III species shared genes coding for complete pathways for the aerobic and anaerobic biosynthesis of cobalamin (Fig. 4*B*). The presence of this metabolic pathway was noted in all species with atypically large genomes, while it is missing in the few Clade III species possessing genome sizes more typical for *Enterococcus* species. Importantly, these genes were not found outside of Clade III, and an in vitro growth experiment using a defined medium, confirmed their role in the de novo biosynthesis of the otherwise essential vitamin B12 (Fig. 4*C*).

Finally, species of Clade IV were enriched in only two orthologous gene groups of largely unknown function. However, a large subset of Clade IV species (>40%) encoded genes necessary for the biosynthesis of flagella, which were otherwise missing in any other species throughout the genus (Fig. 4*D*). This cluster, first described in *E. gallinarum* and *E. casseliflavus* (20), shares identity with the flagellar system in lactobacilli and a similar system (at the gene content level) was also detected in >50% of the known *Vagococcus* species (*SI Appendix*, Fig. S2). Clade IV species are depleted in only 1 orthologous group compared to all other clades (Dataset S3).

Functional Analysis of Novel Species. A comparison of gene content was made between the 18 novel *Enterococcus* species identified here and their nearest taxonomic neighbor (Table 1), to gain insight into what may have selected for speciation. Nearest neighbors were defined as the previously known species that shared the highest genome-wide ANI and sharing the shortest branch length (Fig. 3*A* and Table 1). Where multiple reference genomes existed, gene content was required to be present or absent in all genomes of that species. Genes were first binned into functional categories using COG annotation (*SI Appendix*, Fig. S3*A* and Dataset S6). As found in our prior study (2), carbohydrate transport and metabolism was the largest functional category enriched in species-specific gene content (*SI Appendix*, Fig. S3*A*). Following this, the second most enriched category was suspected defense mechanisms (8.8%), with most encoding ABC-type transporters predicted to efflux antimicrobials and other noxious compounds. Regulation of transcription (7.6%), and signaling (6.9%), largely associated with responding to sources of nutrition, lend further support to the co-evolution of enterococci along with the dietary habits of their hosts (2). Descriptions of each novel species, following SeqCode (21) nomenclatural code are available in *SI Appendix* and Dataset S2.

To better understand the unique biology of the newly identified species, we identified genes likely to be core to novel species but not present in the core or auxiliary genes of other species (i.e., unique in the pan-genome). For this, genes sharing <65% amino acid sequence identity over the full-length inferred protein sequence with genes occurring in any enterococcal genome (a cutoff exceeding the ANI distance between the most divergent enterococcal species), and which did not occur in proximity to known mobile elements, were identified as likely unique and core to each novel species (Dataset S6). A total of 1,276 such genes were identified as potentially species-specifying (Dataset S6), expanding the pan-genome of our sample set of enterococci by 9%. Although many novel genes were of unknown function, we were able to assign a functional annotation to 62% using a combination of annotation and structure-based prediction. This revealed several notable functions encoded by novel species (Table 1 and Dataset S6).

Six additional members of Clade I were identified, including candidate species *Enterococcus* sp. nov. *clewellii* isolated from a captive African Bullet cockroach (Table 1). This species forms an early branch within Clade I in a lineage that gave rise to all Clade I species except for *E. faecalis*. The closest known species to *E. clewellii* is *E. quebecensis*, although the long branch length separating these two species suggests that *E. clewellii* is much more distantly related to *E. quebecensis* than are other novel species and their nearest known relatives. The second novel Clade I species, candidate species *Enterococcus dunnyi* was isolated twice in our sampling: once from a ground beetle, and once from a captive African Bullet cockroach (Table 1). Its closest taxonomic relative, *E. termitis*, is known to colonize the guts of termites (22). The third novel Clade I species, candidate *Enterococcus mansonii*, was isolated from the droppings of a turkey. Its closest taxonomic relative, *E. quebecensis*, was originally isolated from contaminated water samples (23). The fourth novel Clade I species candidate *Enterococcus ikei*, most closely related to *E. ureilyticus*, was isolated twice in our collection from a captive moth and a wild dragonfly. The fifth novel Clade I species candidate, *Enterococcus lemimoniae*, isolated from a moth and two dragonflies, was found to be a very close taxonomic relative of *E. rotai*. *E. rotai* has previously been isolated from mosquitos, water, and plants (24). The sixth novel species added to Clade I was candidate *Enterococcus palustris*, isolated from a dragonfly, which branches ancestrally to *E. haemoperoxidus* (Table 1).

Four more members of Clade II were identified. These include candidate *Enterococcus huntleyi*, which was isolated from chicken feces and is most closely related to *E. phoeniculicola* (80.2% ANI; Table 1). The second novel Clade II species, candidate *E. courvalinii*, was isolated from a cockroach and from turkey droppings, and was most closely related to *E. thailandicus*. The third novel species of Clade II, candidate *Enterococcus wittei*, was isolated from chicken droppings. It was one of two novel species closely related to *E. mundtii*. The fourth novel species of Clade II, also closely related to *E. mundtii*, candidate species *E. mangumiae*, was isolated from two butterflies and one dragonfly in our collection (Table 1).

In Clade III, three previously undescribed species were identified. Interestingly, all three novel species were isolated from Kemp's ridley sea turtles that had stranded on the coast of Massachusetts due to cold-stunning. These species included candidate *E. moelleringii*, which appears to branch early on from the rest of the Clade III strains and is most closely related to *E. pallens*. The second novel species of Clade III was also most closely related to *E. pallens*. This species, candidate *E. ferrettii*, branches more recently from *E. pallens*. The third novel species, candidate *Enterococcus murrayae*, was most closely related to *E. hulanensis*, which is itself a recently discovered species of *Enterococcus* that was isolated from Chinese traditional pickle juice (25).

Within Clade IV, we identified five novel species. These include candidate *E. leclercqii*, isolated from chicken droppings, which is most closely related to *E. diestrammenae*. The second novel species identified in Clade IV, *E. myersii*, was isolated from aquarium water. It branched ancestrally to species *E. dispar* and *E. canintestini*. Candidate *E. testudinis*, isolated from a sea turtle, was intermediate to *E. casseliflavus /E. flavescens* and *E. gallinarum*. Its closest taxonomic neighbors are the *E. casseliflavus/E. flavescens* group. A fourth novel Clade IV species, candidate *Enterococcus lowellii*, isolated from a captive cockatoo, also grouped with *E. saccharolyticus* to further define the taxonomic group in Clade IV that branched most recently from the modern-day tetragenococci. A fifth Clade IV species, candidate *Enterococcus willemsii* isolated from chicken droppings, grouped into a radiation with *E. saccharolyticus* and *E. lowellii*, forming a cluster of species most closely related to the tetragenococci (Fig. 3*A*).

Discussion

As enterococci occur in the guts of the diversity of land animals, from invertebrates to mammals (2, 4, 5), host–microbe interactions can be examined and compared in an unusually wide variety of ecological contexts and physiological systems. This provides an opportunity to identify universal principles that govern host/gut microbe interactions. Enterococci also have emerged among leading causes of antibiotic-resistant hospital-associated infection (1, 26), making understanding of host association principles imperative. Since the 1960s and 1970s work of Mundt and colleagues (4–6, 27), when fewer than 10 enterococcal species had been defined using phenotypic criteria; now over 60 species have been proposed (28) with most associated with at least one representative genome sequence. With no exception, known enterococcal species all differ from their closest sister species by >5% ANI, the threshold proposed as a new "gold standard" for species definition (18). Here, we described the isolation and distribution of many known species from diverse hosts and geographies and further identified 18 previously unknown lineages that qualify as new species by the above criterion (Table 1).

The isolation of 18 previously undescribed species from diverse wild samples suggests that much species-level diversity within the genus *Enterococcus* remains to be discovered. We used our isolation data, and published estimates of global animal biomass and animal species diversity (29–31), to understand the depth to which we had sampled the diversity of different animal taxa, and to extrapolate an upper bound for enterococcal species diversity (Fig. 5*A*). Compared to the total biomass of terrestrial arthropods, there is far less biomass of wild birds (3%) and mammals (1%). We speculate that our recovery of limited enterococcal species variation in birds and mammals (with the sole exception of insectivorous fowl), and wider species variation in arthropods, is a reflection of the disparity in both global biomass as well as limited physiological variation within the vertebrates. A comparison of *Enterococcus* species diversity to the fraction of animal group diversity sampled highlights the depth to which each group was sampled (Fig. 5*B*). This analysis revealed that mammalian diversity was well sampled with few new enterococcal species arising but that the tremendous diversity of terrestrial arthropod hosts was particularly fertile with <0.01% of total arthropod host species diversity having been explored. There are only ~10^3 existing species of mammals, all with distal gut physiologies that are broadly similar. In stark contrast, there are estimated to be 10^6 to 10^7 species of arthropods with widely varying diets [carnivorous (32) to highly specialized herbivores (33)] and gut physiologies (ranging from alkaline to acidic) (34). This study found most of the new species either in

the comparatively few insect and invertebrate samples examined (enterococci were isolated from 140 total terrestrial Arthropod samples; see Dataset S1), or in insectivorous agricultural poultry that are likely capable of serving as aggregators of insect-associated enterococcal species. Based on the small sample size and limited diversity of insects and invertebrate samples examined, we project the existence of thousands of undiscovered species of enterococci defined as varying from the closest extant relative by >5% ANI. Furthermore, additional sampling of sparsely covered geographic regions in our dataset could also increase observed diversity, since our collection was heavily skewed toward strains isolated from hosts occurring in temperate zones in the Northern Hemisphere. Further, we observed a paucity of mobile elements in the new species identified, highlighting that our understanding of the nature of enterococci is heavily skewed by the vast majority of enterococcal genome sequences being derived from strains recovered from sites of infection in hospitalized patients—isolates from habitats where enterococci naturally do not occur but now are actively adapting to.

In this work, as did Mundt and colleagues before us (4–6, 27), we observed several enterococcal species (e.g., *E. faecalis*, *E. mundtii*, and *E. faecium*) associated with a wide variety of hosts, whereas others were rarer, potentially as a consequence of being host specific (Fig. 2*A*). With the widest distribution of all, we deduce that *E. faecalis* is likely the most generalist of species able to proliferate within diverse hosts and survive outside of the host, two factors critical for wide transmission and occurrence. In contrast, the plethora of rarer species suggests that many enterococci have evolved more specialist host associations that favor specific hosts in restricted environments. The paths through which generalist and specialist enterococcal species evolve remain unknown and require the examination of larger collections with specific hosts represented multiple times to determine tropisms.

Hosts themselves vary widely in which *Enterococcus* species they harbor. In contrast to known specific associations, such as that of *E. columbae* with the Columbidae family of pigeons and doves as previously discussed (2), humans are typically colonized by either *E. faecalis* or *E. faecium*, or both (35). In contrast, chicken and poultry are colonized by a wide diversity of species, as observed here and previously (36). In the present study in addition to 10 known species, we identified 6 new species of *Enterococcus* associated with poultry (Table 1). It seems unlikely that these 6 new species (and as many more established species of *Enterococcus* observed by us and others) each evolved specifically to colonize the poultry gut. In nature, poultry and closely ancestrally related birds are highly adapted foragers that exhibit an innate ground-scratching behavior, unearthing insects, and other sources of nutrition (37). It is plausible that early exposure to microbes in the diet shapes the assembly of gut communities in this context and that opportunities exist for the transient occurrence of specialist and generalist enterococci to be isolated from poultry feces. A similar phenomenon may also underlie the diversity of enterococcal species that we observed in dragonflies (Dataset S1). We isolated 10 *Enterococcus* species, including five previously unknown species, from this carnivorous order of insects. Dragonflies eat other insects, and the microbiota of their prey has been found to shape their gut microbiota composition (38). Together, these results suggest that animals whose diet contains a large proportion of insects and other arthropods may be aggregators and reservoirs of particularly diverse enterococci.

The acquisition of adaptive genes can facilitate genetic differentiation by allowing individuals carrying these genes to proliferate in new ecological niches (39). Our finding that many Clade III strains have acquired vitamin B12 biosynthesis pathways suggests that the ability to make this vitamin may be a key trait for the ecology that Clade III strains inhabit (Fig. 4*B*). As we showed, these pathways in fact relieve the B12 auxotrophy that is otherwise common to all other enterococci (Fig. 4*C*). B12 is rare in diets derived purely from plant or fungal matter (40). Cobalamin is made and shared among microbes in various complex consortia (41) including those of the gut. All 3 phylogenetically separate Clade III species newly identified in this study derived from cold-stunned Kemp's ridley sea turtles that came ashore on Cape Cod. These turtles often live in shallow coastal marine water and consume a diet of mainly crustaceans, and to a lesser extent plants (42). The occurrence of three separate Clade III species in samples of the same host type from similar locations suggests that they may derive from the diverse coastal crustaceans consumed. An alternative explanation would be that these species are each independently adapted to the Kemp's ridley turtle; however, this seems less likely in a shared ecosystem where one host-adapted species would likely predominate over time. Many crustaceans subsist on a diet of algae, photosynthetic eukaryotes at the base of the marine food chain that are incapable of synthesizing cobalamin (40). A cobalamin-synthesizing *Enterococcus* in the guts of coastal algae consumers would seem well placed. Together with the intriguingly large genomes of species within this clade, these data suggest that Clade III enterococci occupy a novel ecological niche compared to other species within the genus.

The prevalence of amino acid auxotrophy including branched-chain amino acids and histidine (Fig. 4*E*) is a notable signature of clade II enterococci. For members of the closely related genus *Lactococcus*, auxotrophy for branched-chain amino acids and histidine, which are abundant in milk protein, is believed to be a hallmark of mammalian adaptation (43, 44), a finding that has been recapitulated under laboratory evolution conditions (45). Our previously proposed timeline for *Enterococcus* evolution positions the radiation that split Clade II from Clades III and IV about 350 mya (2), which would predate the evolution of mammals from their synapsid ancestors [approximately 200 mya (46)]. This suggests that the radiation of Clade II may have resulted from an earlier adaptation by enterococci to the guts of the synapsid precursors of mammals, animals that dominated terrestrial life beginning approximately 318 mya (47). The disappearance of the last surviving non-mammalian synapsid about 30 mya (48) now leaves only mammalian representatives of that division—and their gut microbes—to be studied, potentially explaining why surviving Clade II enterococcal species share amino acid auxotrophies characteristic of mammalian-adapted lactococci. Of clinical relevance, diets that include lactose have been observed to predispose hospitalized patients to *E. faecium* overgrowth and subsequent infection (49). It is notable that *E. faecalis*, uniquely among Clade I species, showed similar patterns of amino acid auxotrophy. Future work will be needed to determine ecological and evolutionary factors that led to these convergent patterns of auxotrophy in the two species of *Enterococcus* in which clinically relevant multidrug resistance has emerged.

Our study identified important genetic novelty in diverse enterococci, including a wealth of genes of unknown function yet to be explored. From among these genes, we previously reported the identification of an entirely new class of botulinum-type toxin encoded on a plasmid that occurred in an *E. faecium* excreted from the gut of a cow (10), as well as a new class of pore-forming toxins distantly related to the delta toxins of clostridia, in domesticated animal-derived strains of *E. faecalis, E. faecium*, and *E. hirae* (11), the three most commonly encountered generalists in this study, which have implications for human health. Expanding this search now in a more informed and directed way will illuminate the scope of enterococcal

genetic diversity and potential sources and routes of transmission of antibiotic resistance genes, providing insight into the mechanistic basis for host association of this important class of microbes.

Methods

Isolation of a Library of Diverse Enterococci. Approximately one gram of fecal sample, small whole insects, or gut tracts dissected from large invertebrate samples were mechanically disrupted using a mortar and pestle in 1 to 2 volumes of sterile pH 7.4 phosphate-buffered saline. Sample dissections were carried out with sterilized equipment. Glycerol was added to homogenate as a cryoprotectant, and remaining material was stored at −80 °C. Dilutions of the homogenate were directly plated onto bile esculin agar (BEA; Difco 299068). This medium contains oxgall bile as a selective agent and produces a differential colorimetric reaction upon esculin hydrolysis. Bile resistance is a trait that defines the genus *Enterococcus* (26), and many species hydrolyze esculin. Enterococci were enriched by plating a 100 μL aliquot of the homogenate onto a 25 mm #1 Whatman filter paper placed on a BEA agar plate. The filter papers were incubated at room temperature for 48 to 72 h, or until black pigment developed. Bacterial growth was collected by rinsing filters in brain–heart infusion (BHI) medium (Difco DF0418) and plated onto BEA agar to retrieve single colonies. Morphologically distinct isolated colonies were selected and purified by a further plating onto BEA or CHROMagar Orientation medium. Isolates that grew on BEA, or that produced blue pigmented colonies on CHROMagar were selected for taxonomic identification. Isolates were routinely maintained in BHI and frozen at −80 °C in BHI containing 20% glycerol for long-term storage.

Identification of Isolates. As a template for molecular analysis, genomic DNA was collected from cells grown in BHI with a Blood and Tissue kit (Qiagen), following the manufacturer-supplied protocol with the addition of mutanolysin to fully disrupt enterococcal cell walls during cell lysis. Isolates positively identified as *Enterococcus* spp. were selected for further analysis. Orthologs of *E. faecalis* V583 gene EF1948 present in a sample set of 47 species of *Enterococcus* were aligned to identify conserved regions, and oligonucleotide primers were designed to amplify a conserved region flanking 97bp of more variable sequence called DL (*SI Appendix*, Table S1). The annealing temperature of the primers was optimized by gradient PCR using genomic DNA from 24 known enterococcal species, as well as DNA from 8 non-enterococcal controls from the genera *Streptococcus, Lactococcus*, *Vagococcus,* and *Carnobacteria*. An annealing temperature of 51 °C correctly amplified a single product from each enterococcal template but did not amplify products from the non-enterococcal controls. Where the DL primers failed to amplify a product, the full-length 16S rDNA sequence amplified with primers 8F and 1522R (50) with High Fidelity Q5 Taq polymerase (NEB). A high-quality sequence was obtained by the Sanger method and was used to query the 16S rRNA BLAST database. In all instances, isolates that failed to amplify a product were found to be species other than *Enterococcus*.

All isolates with significant novelty, and those sequenced for further analysis, were registered with the appropriate government entities, working with local collaborators, in agreement with the Nagoya protocol.

Genome Sequencing and Assembly. For 33 isolates, genomic DNA was used to generate dual-indexed Nextera-XT short-read sequencing libraries (Illumina). We sequenced these libraries in paired-end format with 250-bp read length. We performed de novo assembly of short reads using CLC Workbench, after trimming adapter sequences, filtering reads based on quality score and removing PhiX. For the remaining 14 isolates, we prepared whole-genome paired-end and mate-pair Illumina libraries and sequenced and assembled them as previously described (2). We also sequenced 10 of the novel isolates using PacBio Sequel. Libraries were prepared using the SMRTbell Express Template Prep kit 1.0. PacBio assemblies were generated using HGAP 4.0 (51) run through the smrttools (v5.1.0) command-line pbsmrtpipe.pipelines.polished_falcon_fat workflow. The parameters used for the genome assembly included the following settings: falcon_ns.task_options.HGAP_AggressiveAsm_bool set to true, falcon_ns.task_options.HGAP_FalconAdvanced_str set to pa_dbsplit_option=-x500 -s200, falcon_ns.task_options.HGAP_GenomeLength_str set to 3500000, and all other parameters used default settings. The assembly was polished by aligning Illumina reads to assembly contigs using bwa mem (v0.7.4) (52) and then Pilon (v1.23) (53) was run with the --fix bases parameter. Both Illumina and PacBio sequencing reads and assemblies were submitted to GenBank under BioProjects PRJNA324269 (54) and PRJNA313452 (55). Novel species names were registered at SeqCode (21).

In Vitro Growth Assays for Enterococcus spp. on Medium with or without Cobalamin. A total of four enterococcal strains (*E. faecalis* V583, *E. gilvus* BAA-350, *E. avium* ATCC14025, and *E. raffinosus* ATCC49464), including three large-genome clade III species which either did (*E. raffinosus* and *E. avium*) or did not (*E. gilvus*) include the cobalamin biosynthesis pathway, were grown on BUG+B agar plates at 30 °C. Cells were harvested and resuspended in 5 mL PBS to reach OD590 = 0.1. Cell suspensions were centrifuged, washed twice, and resuspended in 1:1 PBS. A 1% inoculum from each cell suspension was inoculated in a vitamin B12-free chemically defined medium (CDM) and grown overnight at 37 °C. The vitamin B12-free medium was modified from an existing CDM used for lactic acid bacteria (56) and contained the following per liter: 1 g K_2HPO_4, 5 g KH_2PO_4, 0.6 g ammonium citrate, 1 g acetate, 0.25 g tyrosine, 0.24 g alanine, 0.125 g arginine, 0.42 g aspartic acid, 0.13 g cysteine, 0.5 g glutamic acid, 0.15 g histidine, 0.21 g isoleucine, 0.475 g leucine, 0.44 g lysine, 0.275 phenylalanine, 0.675 g proline, 0.34 g serine, 0.225 g threonine, 0.05 g tryptophan, 0.325 g valine, 0.175 g glycine, 0.125 g methionine, 0.1 g asparagine, 0.2 g glutamine, 10 g glucose, 0.5 g L-ascorbic acid, 35 mg adenine sulfate, 27 mg guanine, 22 mg uracil, 50 mg cystine, 50 mg xanthine, 2.5 mg D-biotin, 1 mg riboflavin, 5 mg pyridoxamine-HCl, 10 μg *p*-aminobenzoic acid, 1 mg pantothenate, 5 mg inosine, 1 mg nicotinic acid, 5 mg orotic acid, 2 mg pyridoxine, 1 mg thiamine, 2.5 mg lipoic acid, 5 mg thymidine, 200 mg $MgCl_2$, 50 mg $CaCl_2$, 16 mg $MnCl_2$, 3 mg $FeCl_3$, 5 mg $FeCl_2$, 5 mg $ZnSO_4$, 2.5 mg $CoSO_4$, and 2.5 mg $CuSO_4$. When necessary, vitamin B12 was supplemented with the addition of 100, 10, or 1 μg cyanocobalamin per 100 mL (CN-Cbl, Sigma USA). Optical density at 590 nm (OD590) was monitored hourly for 24 h, using a Synergy 2 microplate reader (Bio-Tek).

Comparative Genomics.

Selection of Comparator Genomes. We selected known species of *Enterococcus* and other members of the taxonomic family Enterococcaceae that were also found in guts, choosing those with assemblies of at least 96% completeness, as estimated by CheckM v1.1.3 (57). In many cases, a single representative meeting this criterion was available for each species. Where multiple representatives of a species were sequenced, we selected a representative for each. However, because *E. casseliflavus* genomes showed a high degree of intra-specific variation in ANI, we selected assemblies from three strains of *E. casseliflavus* isolated from human, plant, and chicken sources (Dataset S2), as well as an assembly of *E. flavescens*, which has previously been proposed as a species of *Enterococcus* but is very closely related to *E. casseliflavus* (58, 59). We added a total of 33 comparator genomes, resulting in a total of 103 genomes for comparative analysis.

Annotation of Genomes. We uniformly annotated the 47 newly sequenced *Enterococcus* and *Vagococcus* isolates, as well as the set of comparator genomes, using the Broad Institute's prokaryotic annotation pipeline (17). Furthermore, we annotated coding regions using i) BlastKOALA (make_csd v1.0, exec_koala v1.2, make_aalist v1.0, replacing BLAST with DIAMOND v0.9.24.125) to identify genes with KEGG annotations (60, 61); ii) the CARD database (downloaded October 2019) to annotate antibiotic-resistance genes using RGI v5.1.0 (62); iii) dbCAN (using hmmscan-parser.sh from 07/21/2015 with database v7 downloaded) to identify carbohydrate-active enzymes described in the CAZy database (63, 64); iv) AntiSMASH v4.2.0 (run with options: "-c 16 --clusterblast --subclusterblast --knownclusterblast --borderpredict --asf --transatpks_da --smcog --full-hmmer") to identify secondary metabolite gene clusters, including lanthipeptide synthases (65); v) CRISPRDetect v2.2 to identify CRISPR spacers (66); and vi) ProphET v0.5.1 (with default parameters) to identify prophage (67).

Calculation of ANI. To calculate ANI between the whole genome sequences in our dataset, we used FastANI v1.32 (18), which performs a kmer-based comparison.

Clustering of Orthologous Genes and Core Gene Phylogeny. We identified orthologous clusters of genes across our complete set of 103 genomes using OrthoFinder v2.3.3 (with default parameters) (68), which is optimized for highly diverse datasets. In order to generate a phylogenetic tree, we identified the set of 320 orthologous groups representing genes found in single copy across all isolates (i.e., SCC), performed multiple-sequence alignment using MAFFT-linsi v7.407 (69), converted this alignment to a codon-based alignment using PAL2NAL v14 (70), and then used this alignment to construct a phylogenetic tree using IQ-TREE (v1.7-beta9) (71) with 1,000 bootstrap replicates, an edge-proportional partition model, and using ModelFinder Plus to find the best codon model for

each gene. Phylogenies were visualized using iTOL (72, 73). To identify the four deeply branching clades previously identified (2), we examined the presence of strains that overlapped between the two studies.

Clade-Level Enrichment Analysis. We identified traits enriched or depleted in specific enterococcal clades as well as differentiating the entire genus from outgroups by requiring a trait to be present in at least 80% of the members of one clade and at most 20% of the members of the comparing clade. Traits tested for differential presence between taxonomic clades included: KEGG Ortholog Groups, KEGG Modules, KEGG Pathways, OrthoFinder Ortholog Groups, CARD AMR alleles, and CAZy Carbohydrate Utilization Enzymes.

Comparison of Novel Species Gene Content to Nearest Taxonomic Neighbor. The shared gene content of comparator species was calculated by reciprocal BLAST of gene sequences, using a cutoff of 64% identity across full-length nucleotide sequences of annotated genes. This % cutoff was used because the two least related enterococci were found to share 65% average shared nucleotide identity. Genes identified by reciprocal BLAST were identified as being present in both genomes. Where multiple comparator genomes existed, genes had to be present in all comparators to be considered shared, or present only in novel species comparators to be considered part of the set that differentiated the novel species from its nearest taxonomic neighbor. The predicted function of genes present only in novel species was further analyzed by mapping COG annotations to their letter code, which defined larger functional classes.

Analysis of Gene Content Characterizing Novel Species. To investigate the unique characteristics encoded in the novel species, we identified and characterized 1,426 genes that were found in a novel species, and in no other enterococcal species, using our orthogroup clusters. In addition to the functional annotation described in *Annotation of Genomes*, we performed structure-based function prediction for genes of unknown function using PHYRE2 (v2.0) with a confidence or sequence identity threshold of 90% (74). To identify genes which may have originated from horizontal gene transfer, we searched for mobile elements, including Prophages (*Annotation of Genomes*) and IS elements in close proximity (10 genes upstream or downstream) using ISfinder (75). We also identified clusters of these genes which colocalized on the chromosome of a particular novel species, or genes which were at most 10 genes apart on the same scaffold.

Estimating Species Diversity within the Genus Enterococcus. We gathered estimates of total animal biomass, and species diversity within taxonomic groups of animals from the literature (29–31, 76). We estimated the depth to which we had sampled animal species diversity by dividing the total number of samples collected for one animal type by the published estimates of total species. This approximation overestimates the depth of sampling where we sampled multiple animals belonging to the same species. We chose this method because it was not possible to make a species-level taxonomic identification for some sample types such as the scat of wild animals or some whole arthropod samples.

Data, Materials, and Software Availability. Illumina and PacBio sequencing reads and assemblies have been deposited in NCBI Genbank (BioProjects PRJNA324269 (54) and PRJNA313452 (55)).

ACKNOWLEDGMENTS. Samples were collected by the Enterococcal Diversity Consortium, a worldwide consortium of collaborators, including Adventure Scientists, Consumer Reports, the Marine Resources Center, the New England Aquarium, the Clemson University Morgan Poultry Center, the Eisen Laboratory, the Kolter Laboratory, Alexander Bertonneau, Kristal Bertonneau, Sophia Bertonneau, Peter Billman, Allen Bolinger, Robert Bruker, Ilana Camargo, Peter Claussen, Tucker Cunningham, Lonnie Dupre, Colleen Ferris, Nkrumah Frazier, Ana Frazzon, Matt Gaidica, Marla Garrison, Rebecca Gast, Michael Gilmore, Peter Girguis, Gonzalo Giribet, Erin Gontang, Jennie Groves, Sebastian Gunther, Wolfgang Haas, Devin Huntley, Suzie Imber, Charles Innis, Mike Libecki, Abigail Manson, Joseph Manson, Pascale Marceau, Megan May, Nathan McGuire, Patrick McGuire, Richard McLaughlin, Philip Metzger, Joanne Munisteri, Andrew Oster, Janira Prichula, Matthew Rowbottom, Stefani Ryan, Katharina Schaufler, Róza Sebök, Helen Seneker, Bruna Sgardioli, Ken Tennessen, Gregg Treinish, Daria Van Tyne, Hera Vlamakis, Jeff Vohl, Jaap Wagenaar, Maarten Gilbert, Jenna Wallenga, Jeremy Wei, and Sheila Withrow. Those enterococci isolated in Sao Carlos, Brazil, are registered at the National System for the Management of Genetic Heritage and Associated Traditional Knowledge of Brazil, SISGEN, under the number A85F977, and those isolated in Porto Alegre, Brazil, under SISGEN number A720680. This project was supported by the Harvard-wide Program on Antibiotic Resistance NIH/NIAID grant AI083214 and U19AI110818 to the Broad Institute. Portions of the work were supported by a Research Sabbatical grant to M.S.G. from Research to Prevent Blindness to explore the origins of antibiotic resistance. J.A.S. was supported by the NIH Ruth Kirschstein fellowship F32GM121005.

Author affiliations: [a]Department of Ophthalmology, Mass Eye and Ear, Harvard Medical School, Boston, MA 02144; [b]Department of Microbiology, Harvard Medical School, Boston, MA 02115; [c]Department of Biology, University of Southern California, Los Angeles, CA 90089; [d]Multidrug-Resistant Organism Repository and Surveillance Network, Walter Reed Army Institute of Research, Silver Spring, MD 20910; [e]Infectious Disease and Microbiome Program, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142; [f]Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53706; [g]University of Greifswald, Institute of Pharmacy, Greifswald 17489, Germany; [h]Kiel University and University Medical Center Schleswig-Holstein, Institute of Infection Medicine, Kiel 24105, Germany; [i]Delft Bioinformatics Lab, Department of Intelligent Systems, Delft University of Technology, Delft 2628XE, The Netherlands; [j]Laboratório de Epidemiologia e Microbiologia Moleculares, Departamento de Física e Ciências Interdisciplinares, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos - SP 13566-590, Brazil; [k]Federal University of Health Sciences of Porto Alegre, Porto Alegre - RS 90050-170, Brazil; [l]Department of Microbiology, Immunology and Parasitology, Federal University of Rio Grande do Sul, Porto Alegre - RS, 90010-150, Brazil; [m]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138; [n]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburg, PA 15213; [o]Adventure Scientists, Bozeman, MT 59715; [p]New England Aquarium, Animal Health Department and Anderson Cabot Center for Ocean Life, Boston, MA 02110; and [q]Department of Biomolecular Health Sciences, Utrecht University, Utrecht 3584 CS, The Netherlands

Author contributions: J.A.S., F.L., A.L.M., A.M.E., and M.S.G. designed research; J.A.S., F.L., R.S., M.J.M., K.S., B.F.S., R.M.W., and A.L.M. performed research; I.L.B.C.C., J.P., A.P.G.F., G.G., D.V.T., G.T., C.J.I., J.A.W., and M.S.G. contributed new reagents/analytic tools; J.A.S., F.L., R.S., T.S., A.U., T.A., A.L.M., A.M.E., and M.S.G. analyzed data; and J.A.S., F.L., R.S., A.L.M., A.M.E., and M.S.G. wrote the paper.

1. O. Gaca Anthony, A. Lemos José, Adaptation to adversity: The intermingling of stress tolerance and pathogenesis in Enterococci. *Microbiol. Mol. Biol. Rev.* **83**, e00008-19 (2019).
2. F. Lebreton *et al.*, Tracing the Enterococci from Paleozoic origins to the hospital. *Cell* **169**, 849–861. e13 (2017).
3. F. Lebreton, R. J. L. Willems, M. S. Gilmore, "Enterococcus diversity, origins in nature, and gut colonization" in Enterococci: From Commensals to Leading Causes of Drug Resistant Infection, M. S. Gilmore, D. B. Clewell, Y. Ike, N. Shankar, Eds. (Massachusetts Eye and Ear Infirmary, 2014).
4. J. O. Mundt, Occurrence of Enterococci in animals in a wild environment. *Appl. Microbiol.* **11**, 136–140 (1963).
5. J. D. Martin, J. O. Mundt, Enterococci in insects. *Appl. Microbiol.* **24**, 575–580 (1972).
6. J. O. Mundt, A. H. Johnson, R. Khatchikian, Incidence and nature of Enterococci on plant materials. *Food Res.* **23**, 186–193 (1958).
7. J. O. Mundt, Occurrence of Enterococci: Bud, blossom, and soil studies. *Appl. Microbiol.* **9**, 541–544 (1961).
8. K. H. Schleifer, R. Kilpper-Balz, Transfer of *Streptococcus faecalis* and *Streptococcus faecium* to the Genus *Enterococcus* Nom. Rev. as *Enterococcus faecalis* Comb. Nov. and *Enterococcus faecium* Comb. Nov. *Int. J. Syst. Bacteriol.* **34**, 31–34 (1984).
9. R. R. Facklam, Comparison of several laboratory media for presumptive identification of Enterococci and Group D Streptococci. *Appl. Microbiol.* **26**, 138–145 (1973).
10. S. Zhang *et al.*, Identification of a *Botulinum* neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*. *Cell Host Microbe* **23**, 169–176.e6 (2018).
11. X. Xiong *et al.*, Emerging Enterococcus pore-forming toxins with MHC/HLA-I as receptors. *Cell* **185**, 1157–1171.e22 (2022).
12. J. Prichula *et al.*, Corrigendum "Resistance to antimicrobial agents among enterococci isolated from fecal samples of wild marine species in the southern coast of Brazil" [Mar. Pollut. Bull. 105 (2016) 51–57]. *Mar. Pollut. Bull.* **141**, 655–656 (2019).
13. J. Prichula *et al.*, Enterococci from wild magellanic penguins (*Spheniscus magellanicus*) as an indicator of marine ecosystem health and human impact. *Appl. Environ. Microbiol.* **86**, e01662–20 (2020).
14. J. Merlino *et al.*, Evaluation of CHROMagar orientation for differentiation and presumptive identification of gram-negative bacilli and Enterococcus species. *J. Clin. Microbiol.* **34**, 1788–1793 (1996).
15. L. A. Devriese *et al.*, Differentiation and identification of *Enterococcus durans*, *E. hirae* and *E. villorum*. *J. Appl. Microbiol.* **92**, 821–827 (2002).
16. M. V. Belloso Daza, C. Cortimiglia, D. Bassi, P. S. Cocconcelli, Genome-based studies indicate that the *Enterococcus faecium* Clade B strains belong to *Enterococcus lactis* species and lack of the hospital infection associated markers. *Int. J. Syst. Evol. Microbiol.* **71**, 004948 (2021).
17. F. Lebreton *et al.*, Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. *mBio* **4**, e00534–13 (2013).
18. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
19. L. Bailey, M. D. Collins, Reclassification of "*Streptococcus pluton*" (White) in a new genus *Melissococcus*, as *Melissococcus pluton* nom. *J. Appl. Bacteriol.* **53**, 215–217 (1982).

20. K. L. Palmer *et al.*, Comparative genomics of *Enterococci*: Variation in *Enterococcus faecalis*, clade structure in *E. faecium*, and defining characteristics of *E. gallinarum* and *E. casseliflavus*. *mBio* **3**, e00318-11 (2012).
21. B. P. Hedlund *et al.*, SeqCode: A nomenclatural code for prokaryotes described from sequence data. *Nat Microbiol* **7**, 1702–1708 (2022).
22. P. Švec *et al.*, *Enterococcus silesiacus* sp. nov. and *Enterococcus termitis* sp. nov. *Int. J. Syst. Evol. Microbiol.* **56**, 577–581 (2006).
23. V. Sistek *et al.*, *Enterococcus ureasiticus* sp. nov. and *Enterococcus quebecensis* sp. nov., isolated from water. *Int. J. Syst. Evol. Microbiol.* **62**, 1314–1320 (2012).
24. I. Sedláček *et al.*, *Enterococcus ureilyticus* sp. nov. and *Enterococcus rotai* sp. nov., two urease-producing Enterococci from the environment. *Int. J. Syst. Evol. Microbiol.* **63**, 502–510 (2013).
25. Y. Q. Li, C. T. Gu, *Enterococcus pingfangensis* sp. nov., *Enterococcus dongliensis* sp. nov., *Enterococcus hulanensis* sp. nov., *Enterococcus nangangensis* sp. nov. and *Enterococcus songbeiensis* sp. nov., isolated from Chinese traditional pickle juice. *Int. J. Syst. Evol. Microbiol.* **69**, 3191–3201 (2019).
26. E. Fiore, D. Van Tyne, M. S. Gilmore, Pathogenicity of Enterococci. *Microbiol. Spectr.* **7**, 1–23 (2019).
27. J. O. Mundt, Occurrence of Enterococci on plants in a wild environment. *Appl. Microbiol.* **11**, 141–144 (1963).
28. D. H. Parks *et al.*, A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
29. Y. M. Bar-On, R. Phillips, R. Milo, The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6506–6511 (2018).
30. G. F. Barrowclough, J. Cracraft, J. Klicka, R. M. Zink, How many kinds of birds are there and why does it matter? *PLoS ONE* **11**, e0166307 (2016).
31. N. E. Stork, How many species of insects and other terrestrial arthropods are there on earth? *Annu. Rev. Entomol.* **63**, 31–45 (2018).
32. M. Coll, M. Guershon, Omnivory in terrestrial arthropods: Mixing plant and prey diets. *Annu. Rev. Entomol.* **47**, 267–297 (2002).
33. M. L. Forister *et al.*, The global distribution of diet breadth in insect herbivores. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 442–447 (2015).
34. J. F. Harrison, Insect acid-base physiology. *Annu. Rev. Entomol.* **46**, 221–250 (2001).
35. D. Krista, E. G. Pamer, B. R. Allen, P. D. Cani, Enterococci and their interactions with the intestinal microbiome. *Microbiol. Spectr.* **5**, 1–16 (2017).
36. M. A. Rehman *et al.*, Genotypes and phenotypes of Enterococci isolated from broiler chickens. *Front. Sustainable Food Syst.* **2**, 83 (2018).
37. K. C. Klasing, Poultry nutrition: A comparative approach. *J. Appl. Poult. Res.* **14**, 426–436 (2005).
38. R. Deb, A. Nair, D. Agashe, Host dietary specialization and neutral assembly shape gut bacterial communities of wild dragonflies. *PeerJ* **7**, e8058 (2019).
39. M. F. Polz, E. J. Alm, W. P. Hanage, Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **29**, 170–175 (2013).
40. J. R. Roth, J. G. Lawrence, T. A. Bobik, Cobalamin (coenzyme B12): Synthesis and biological significance. *Annu. Rev. Microbiol.* **50**, 137–181 (1996).
41. P. H. Degnan, M. E. Taga, A. L. Goodman, Vitamin B12 as a modulator of gut microbial ecology. *Cell Metab.* **20**, 769–778 (2014).
42. E. E. Seney, J. A. Musick, Diet analysis of Kemp's ridley sea turtles (*Lepidochelys kempii*) in Virginia. *Chelonian Conserv. Biol.* **4**, 864–871 (2005).
43. C. Delorme, J. J. Godon, S. D. Ehrlich, P. Renault, Gene inactivation in *Lactococcus lactis*: Histidine biosynthesis. *J. Bacteriol.* **175**, 4391–4399 (1993).
44. J. J. Godon *et al.*, Gene inactivation in *Lactococcus lactis*: Branched-chain amino acid biosynthesis. *J. Bacteriol.* **175**, 4383–4390 (1993).
45. H. Bachmann, M. J. C. Starrenburg, D. Molenaar, M. Kleerebezem, J. E. T. van Hylckama Vlieg, Microbial domestication signatures of *Lactococcus lactis* can be reproduced by experimental evolution. *Genome Res.* **22**, 115–124 (2012).
46. S. Álvarez-Carretero *et al.*, A species-level timeline of mammal evolution integrating phylogenomic data. *Nature* **602**, 263–267 (2022).
47. N. Brocklehurst, The first age of reptiles? Comparing reptile and synapsid diversity, and the influence of Lagerstätten, during the carboniferous and early Permian. *Front. Ecol. Evol.* **9**, 669765 (2021).
48. H. Matsuoka, N. Kusuhashi, I. J. Corfe, A new Early Cretaceous tritylodontid (Synapsida, Cynodontia, Mammaliamorpha) from the Kuwajima Formation (Tetori Group) of central Japan. *J. Vert. Paleontol.* **36**, e1112289 (2016).
49. C. K. Stein-Thoeringer *et al.*, Lactose drives Enterococcus expansion to promote graft-versus-host disease. *Science* **366**, 1143–1149 (2019).
50. S. Turner, K. M. Pryer, V. P. Miao, J. D. Palmer, Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J. Eukaryot. Microbiol.* **46**, 327–338 (1999).
51. C.-S. Chin *et al.*, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
52. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. B. J. Walker *et al.*, Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
54. Massachusetts Eye and Ear Infirmary/Harvard Medical School. Data from Bioproject PRJNA324269. NCBI Genbank. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA324269. Deposited 27 January 2020.
55. Broad Institute. Data from Bioproject PRJNA313452. NCBI Genbank. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA313452. Deposited 16 May 2017.
56. T. Fiedler *et al.*, Characterization of three lactic acid bacteria and their isogenic ldh deletion mutants shows optimization for YATP (cell mass produced per mole of ATP) at their physiological pHs. *Appl. Environ. Microbiol.* **77**, 612–617 (2011).
57. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
58. S. M. Naser *et al.*, Reclassification of *Enterococcus flavescens* Pompei et al. 1992 as a later synonym of *Enterococcus casseliflavus* (ex Vaughan et al. 1979) Collins et al. 1984 and *Enterococcus saccharominimus* Vancanneyt et al. 2004 as a later synonym of *Enterococcus italicus* Fortina et al. 2004. *Int. J. Syst. Evol. Microbiol.* **56**, 413–416 (2006).
59. R. Pompei *et al.*, *Enterococcus flavescens* sp. nov., a new species of Enterococci of clinical origin. *Int. J. Syst. Bacteriol.* **42**, 365–369 (1992).
60. M. Kanehisa, Y. Sato, K. Morishima, BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
61. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
62. B. Jia *et al.*, CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
63. B. L. Cantarel *et al.*, The carbohydrate-active EnZymes database (CAZy): An expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).
64. H. Zhang *et al.*, dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
65. M. H. Medema *et al.*, antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).
66. A. Biswas, R. H. J. Staals, S. E. Morales, P. C. Fineran, C. M. Brown, CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
67. J. L. Reis-Cunha, D. C. Bartholomeu, A. L. Manson, A. M. Earl, G. C. Cerqueira, ProphET, prophage estimation tool: A stand-alone prophage sequence prediction tool with self-updating reference database. *PLoS ONE* **14**, e0223364 (2019).
68. D. M. Emms, S. Kelly, OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
69. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
70. M. Suyama, D. Torrents, P. Bork, PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
71. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
72. I. Letunic, P. Bork, Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
73. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
74. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
75. P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, M. Chandler, ISfinder: The reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
76. C. J. Burgin, J. P. Colella, P. L. Kahn, N. S. Upham, How many species of mammals are there? *J. Mammal.* **99**, 1–14 (2018).