Macrogenetic atlas of prokaryotes reveals selection-driven structures

Chao Yang[1,2*], Fuxing Jiao[1,2,3], Naike Wang[1,2,3], Shanwei Tong[1,2,3], Hao Huang[1,2,3], Wenxuan Xu[1,2,3], Xavier Didelot[4], Ruifu Yang[5], Yujun Cui[5], Daniel Falush[1,2*]

[1] Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China;
[2] Shanghai Institute of Immunity and Infection, Chinese Academy of Sciences, Shanghai, China;
[3] University of Chinese Academy of Sciences, Beijing, China.
[4] School of Life Sciences and Department of Statistics, University of Warwick, UK
[5] State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences, Beijing, China.

* Corresponding authors:
C.Y.: cyang@siii.cas.cn; D.F.: daniel.falush@siii.cas.cn

Abstract

Macrogenetics investigates patterns and predictors of intraspecific genetic variation across diverse taxa, offering a framework for understanding species diversity and addressing evolutionary hypotheses. Here, we present a macrogenetic atlas of prokaryotes (MAP), integrating genomic data (30 parameters in 12 categories) from 15,235 prokaryotic species and population genetic data (22 parameters in 7 categories) from 786 species with phylogenetic, phenotypic, and ecological information. MAP enables quantitative species characterization, straightforward cross-species comparisons, and easy generation, testing, and refinement of evolutionary hypotheses. For example, our analyses show that long- and short-range genetic linkage capture distinct evolutionary dynamics and independently shape genetic diversity, contradicting neutral theory. Instead, we propose that as within-species diversity increases, epistatic selection becomes an increasingly strong force structuring diversity, exemplified by convergent ecospecies structures in *Streptococcus mitis* and *S. oralis*. Overall, MAP represents a widely applicable resource (www.genomap.cn) and offers novel insights into the drivers of macroevolution and the life cycle of prokaryotic species.

Introduction

Life is diverse, and its understanding requires macro-scale frameworks that capture species diversity, akin to maps guiding exploration. Taxonomy has provided this foundation through efforts such as the Tree of Life[1]. Building on this, comparative studies linking genotypes to phenotypes and ecology are essential for characterizing species and identifying the forces driving diversification and speciation. Classical resources such as *Bergey's Manual*, which integrates taxonomy with phenotypic and ecological information[2], have long served this purpose.

Incorporating genomic data is the next critical step. Advances in high-throughput sequencing have produced extensive genomic resources across many taxa[3,4]. With their rapid accumulation, the time is now ripe for systematic integration under a new framework, macrogenetics, which embraces broad taxonomic, spatial and temporal scales across dozens to thousands of species[5,6].

Macrogenetics provides a framework for exploring species diversity and testing long-standing evolutionary hypotheses. For example, community-level species richness has been shown to correlate positively with within-population genetic diversity in fish and mammals[7,8], supporting the species-genetic diversity correlation (SGDC) concept[9]. Other studies have linked genetic diversity to life-history traits such as fecundity and propagule size[10], and revealed latitudinal gradients in genetic diversity[11].

Prokaryotes make up more biomass than all organisms except plants[12]. Furthermore, a single species often contains lineages adapted to multiple niches, differing in host range, pathogenicity and nutrient use, as exemplified by *Escherichia coli*[13]. This contrasts with eukaryotes, where within-species differentiation and specialization is typically limited. Such pervasive intraspecific diversity in prokaryotes raises fundamental questions: how are ecological niches partitioned, and how do phenotypic and niche diversity shape genomic and population genetic features? These differences also imply that differentiation and speciation in prokaryotes is driven by forces distinct from those shaping eukaryotes[14], but these remain poorly understood.

Here we present the Macrogenetic Atlas of Prokaryotes (MAP), an open resource (www.genomap.cn) that integrates large-scale genomic and population genetic data with phylogenetic, phenotypic and ecological information. MAP assigns each species a quantitative "position" within prokaryotic diversity, analogous to latitude and longitude on a map, providing a framework for exploring and understanding species diversity.

Results

Overview of MAP

We analyzed 317,542 assembled genomes across 15,235 species (503 archaeal and 14,732 bacterial; NCBI taxonomy). After deduplication and quality control, we generated representative (*Rep.*) datasets for each species to reduce repeat sampling biases, and clonal group (CG) datasets to capture microevolution (**Fig. 1a**; Methods; Supplementary Tables 1-2, Supplementary Fig.1).

Species-level genomic parameters (GPs) were defined as the median across *Rep.* genomes, whereas population genetic parameters (PGPs) were estimated from random sampling of ten *Rep.*

genomes, from CG datasets, or from all *Rep.* genomes (Methods). To ensure robustness, only PGPs with low variability (standard deviation/median < 0.5) across random sampling replicates were retained, yielding estimates highly consistent with full datasets (correlation coefficient r > 0.8 for all but one parameter; *h/m*, r = 0.68) and independent of sample size (Supplementary Figs. 2-4). By contrast, all GP and CG-based PGP estimates were retained across species to maximize cross-species comparability, with high variability reflecting substantial within-species heterogeneity.

In total, we obtained 30 GPs across 12 categories for all 15,235 species, and 22 PGPs across 7 categories for 786 species. These were integrated with phylogenetic, phenotypic and ecological data (Supplementary Table 3-6). Metadata and key process outputs, including whole-genome alignments and variation profiles, are available as community resources. We illustrate the utility of MAP across five distinct aspects.

Quantitative chracterization of prokaryotic species

First, MAP translates qualitative descriptions into quantitative, comparable measures (**Fig. 1b,c**). For example, it confirms that the human pathogen *Helicobacter pylori* is highly recombinogenic[15], with recombination-derived diversity ('recombination coverage', *Rec. c*[16]) in the top 1% and short-range (10 bp) linkage disequilibrium (LD, inversely correlated with recombination) in the bottom 1%. This qualitative characterization also revises prevailing views. For example, the widely distributed *E. coli*, naturally presumed to have a large effective population size (N_e) and high diversity[13], instead shows moderately high but not extreme N_e (>71% species; estimated from *dN/dS*, a negative proxy for N_e and selection efficacy in many microbial studies[17]) and intermediate genetic diversity (>39% species).

Defining boundaries of prokaryotic diversity

Second, MAP delineates genomic boundaries across prokaryotes (**Fig. 1b-d**). For example, prokaryotes are are generally small and densely coded[18], with MAP revealing a median genome size of 4.0 Mb (interquartile range, IQR: 2.9–5.3 Mb) and coding proportion of 88% (IQR: 86–90%). Genomes larger than 7.3 Mb (top 10%), such as *Burkholderia cenocepacia* with multipartite chromosomes[19], and smaller than 2.1 Mb, such as *H. pylori* (1.6 Mb, bottom 4%), define the upper and lower extremes.

This quantitative framework enables systematic identification of "extreme" species. For example, *H. pylori* carries an unusually high proportion of restriction-modification (RM) system genes (5.2% of the genome, top 1%), potentially reflecting their role in maintaining genome integrity while the species is naturally competent for foreign DNA uptake[20]. *Bordetella pertussis* has an exceptionally high proportion of insertion sequences (IS, 7.3% of the genome, top 1%), possibly providing alternative evolutionary routes via IS-mediated rearrangements[21] in this otherwise extremely low-diversity species (bottom 1%). Other extremes, which seem to be unreported, include highly fluid genomes in gut-dominant *Bacteroides*, such as *B. uniformis* (fluidity □=0.43, 43% of genes fluidic; top 1%) and *B. fragilis* (top 5%), likely due to frequent exposure to exogenous DNA[22], and extremely low fluidity in intracellular pathogens, such as *Brucella melitensis* (bottom 1%, □=0.01) and *B. abortus* (bottom 2%), likely reflecting their restricted lifestyles[23].

Overall patterns of prokaryotic genomes

Third, MAP reveals overall patterns of prokaryotic genomes (**Fig. 1d**). Archaea and bacteria show broadly similar GP distributions, except that archaeal genes are generally shorter. Most parameters follow roughly normal or right-skewed distributions. In contrast, GC content and codon bias index (CBI) are distinctly bimodal, with GC likely reflecting ancient adaptive legacies[24], and CBI strongly influenced by GC[25]. Bacterial genome size approximates a normal distribution, whereas core- and pan-gene counts are bimodal. The distribution of core-genes likely reflects different outcomes of an evolutionary tradeoff but with the constraint that a minimum number of genes is needed for cellular function.

Easy testing and refining evolutionary hypotheses

Forth, MAP systematically quantified correlations among parameters (r for quantitative traits, η for categorical traits) and assessed their relationships with phylogeny (Pagel's λ), phenotype, and ecology (**Fig. 1e, Fig. 2**). These uniformly derived correlation matrices provide an unprecedented empirical resource for testing and refining evolutionary hypotheses.

Determinants of genome size

Genome size, a key but unresolved trait, is predicted by the "balance" model in prokaryotes[17], reflecting a trade-off between diminishing benefits of additional genes, constrained by selection efficacy, and intrinsic DNA deletion bias. MAP systematically validated this across hundreds of species (**Fig. 2a**). After correcting for phylogeny, genome size negatively correlates with *dN/dS* (*r-adj* = -0.22), the proxy for selection efficacy, and with substitution rate *u* (*r-adj* = -0.26) that linked to deletion bias. Notably, their joint effect (*dN/dS*×*u*) shows an much stronger correlation (*r-adj* = -0.59), providing the strongest empirical support for the balance model so far—an analysis previously inaccessible.

MAP further enables straightforward testing of ecological and phenotypic influences (**Fig. 2a**): species from soil or low-temperature habitats tend to have larger genomes, reflecting habitat and temperature effects[26]. Genome size also negatively correlates with growth rate (*r-adj* = -0.13), indicating replication cost constraints.

Determinants of genetic diversity

The determinants of genetic diversity also remain unresolved. N_e, mutation, and recombination have been proposed as major drivers[27]. While partly tested in eukaryotes, no equivalent dataset existed for prokaryotes because it is challenging to directly measure these factors. The MAP parameters that are most closely related to these proposed drivers are *dN/dS* (*r-adj* = -0.76, inversely reflecting N_e), substitution rate *u* (*r-adj* = 0.13) which measures mutation filtered by selection, and recombination proxy such as the recombination-to-mutation ratio (*r/m*[28], *r-adj* = 0.32). All of these correlations with synonymous diversity π_s are in the expected direction, providing preliminary empirical data supporting their role as drivers in prokaryotes (**Fig. 2b**).

MAP confirmed ecological influences: plant-root-associated species exhibited the highest π_s, consistent with nutrient-rich conditions sustaining large populations, whereas fermented-food

species showed the lowest, reflecting artificial selection and bottlenecks. Diversity also positively correlated with temperature (*r-adj* = 0.24), supporting warmer environments as drivers of genetic diversity and aligning with latitudinal gradients reported in both bacteria[29,30] and eukaryotes[11]. Phenotypic correlates paralleled eukaryotic patterns: just as larger propagule size reduces diversity as life-history strategies[10], prokaryotic cell size negatively correlated with π_s (*r-adj* = -0.14), suggesting a potentially universal cross-domain mechanism.

Underexplored correlations

MAP also revealed previously overlooked patterns (**Fig. 2c**). The negative correlation (*r-adj* = -0.31) between gene length and *dN/dS* suggests that, analogous to genome size, stronger selection favors longer genes. Gene length also correlated positively with coding density (*r-adj* = 0.48), possibly because regulatory non-coding regions remain relatively constant regardless of gene size. Substitution rate *u* was negatively correlated with GC content (*r-adj* = -0.19), consistent with universal AT-biased mutation[31], and increased in warmer environments (*r-adj* = 0.18), indicating temperature-driven acceleration of mutation. MAP also uncovered potentially novel drivers of gene fluidity: RM-gene number correlated strongly with fluidity (*r-adj* = 0.57), implying selection for abundant RM systems under frequent mobile-element invasion, given its role in regulating DNA transfer[32].

Overview of prokaryotic genetic diversity

Finally, we illustrate how MAP can refine theoretical frameworks and generate new hypotheses. We focused on the relationship between LD and genetic diversity, summarized in a two-dimensional triangular plot, since short-range (10 bp) LD is constrained to be lower than long-range (10 kb) LD (**Fig. 3a**). First, nearly the whole triangle is occupied, implying short- and long-range LD vary largely independently. Second, π_s showed a strong negative correlation with short-range LD (*r-adj* = -0.58) but no significant correlation with long-range LD, contradicting the expectation under neutral theory. Instead, for species with similar short-range LD, there is a pronounced positive correlation between long-range LD and π_s (**Fig. 3a**). Therefore, short- and long-range LD reflect different processes and have different drivers.

These qualitative patterns can be explained by the joint effects of species age, reflected in genetic diversity, and barriers to gene flow, reflected in long-range LD (**Fig. 3b**). After a strong bottleneck, species begin with low diversity and high LD. As species accumulate age (indicated by increased π_s) and in the absence of barriers, long-range LD decays rapidly, whereas short-range LD equilibrates more gradually (**Fig. 3b**, top arrows). By contrast, barriers to recombination slow the decay of long-range LD, raising its equilibrium level (**Fig. 3b**, bottom arrows). Although simplified, for example because species-wide bottlenecks can both reduce diversity and elevate LD, thereby making species appear “younger”, this two-dimensional framework provides a useful lens for interpreting species-specific LD and genetic diversity configurations.

An example of a species with low barriers to genetic exchange is *Vibrio parahaemolyticus* (**Fig. 3a,c**), which exhibits low long-range LD but moderate diversity and correspondingly moderate short-range LD. Previous analyses suggested that its diversity remains moderate because population size expanded only recently[33], and if maintained over longer timescales, diversity

would increase considerably, driving the species toward panmixia with high diversity and little LD at any distance. By contrast, a species with well-recognized barriers to genetic exchange is *E. coli* (**Fig. 3a,c**), where recombination between phylogroups is constrained by homology-dependent mechanisms[34], resulting in higher long-range LD than *V. parahaemolyticus.* Another example is *H. pylori*, a species notorious for its exceptionally high rates of recombination during mixed infection[15], which have led to very low short-range LD. However, its geographical structure, due to low rate of spread between human hosts as they dispersed out of Africa[35], also leads to higher long-range LD than in *V. parahaemolyticus.*

Strikingly, although many species resemble *V. parahaemolyticus* in exhibiting low long-range LD, none approach complete panmixia. In our dataset, nearly all species display higher long-range LD than *V. parahaemolyticus* (**Fig. 3a**). The closest candidate is the cockroach symbiont *Blattabacterium cuenoti* (**Fig. 3a,c**), yet only 22 genomes are currently available. Given that many bacteria occupy niches favorable to global dispersal and the maintenance of vast, stable populations over evolutionary timescales, there is therefore a paradoxical absence of panmictic species.

Epistatic selection as a driver of population structure

The paradox can be explained if species with low barriers to exchange systematically evolve stronger barriers as genetic diversity increases. *V. parahaemolyticus* illustrates this process: in addition to weak geographic barriers generating oceanic gene pools[36], further barriers arise from epistatic selection. Specifically, two major ecospecies are highly differentiated at ~1% of the core genome, reflecting strong selection against gene flow in these regions[37,38]. If functional diversification expands over time, epistatic selection could progressively create barriers to exchange across larger genomic regions. *V. parahaemolyticus* would then not evolve into a panmictic species but instead come to resemble the structured, high diversity species that we have observed.

We therefore hypothesize that panmictic bacteria are absent because, in species maintaining large N_e, inevitable functional diversification driven by epistatic selection progressively generates long-range LD. To test this, we examined the genetic basis of population structure across species. Specifically, we used principal component analysis (PCA) to identify the major population structure in each species. We hypothesized that this structure should exhibit stronger signatures of selection in species approaching population genetic equilibrium, where short-range LD is similar to long-range LD. Accordingly, we introduce an LD saturation statistic, defined as the ratio of long-range to short-range LD. As a simple measure of selection, we assessed enrichment of non-synonymous variants among SNPs most strongly associated with the first PC component, as measured by SNP loadings. We expect selection to be evident in highly saturated species.

As predicted by the hypothesis, enrichment of non-synonymous sites correlated strongly with LD saturation among species with high saturation (>0.5, zones 3 and 4 in **Fig. 3a, d**), particularly in species with low short-range LD (zone 4), indicating that selection is most evident in highly recombining species. For example, *Pseudomonas fluorescens*, which exhibits very high genetic diversity and low, nearly saturated LD, showed marked enrichment of non-synonymous sites (N/S

ratio = 2.5) and a highly structured phylogeny (**Fig. 3c,d**). In contrast, species with low saturation (zones 1 and 2) showed no enrichment, consistent with insufficient time for genome-wide signatures of selection to accumulate. Together, these results support a model in which epistatic selection progressively shapes prokaryotic population structure, preventing the emergence of panmictic species.

Convergent ecospecies

Given the evidence that selection has molded genetic structure, we asked whether we could identify species further along the road to functional differentiation than *V. parahaemolyticus*. To this end, we examined species with higher diversity and many genomes (>100 *Rep*. genomes). Two such species are *Streptococcus mitis* and *S. oralis*. The PCA loading distribution for both resembled that seen for previous ecospecies[37,39], implying high differentiation in specific regions and high gene flow elsewhere (**Fig. 4a,b**). Furthermore, there is strong enrichment of genes of particular categories with extreme PC1 loadings. For both species, COG categories D (cell cycle control, cell division, chromosome partitioning) and M (cell wall/membrane/envelope biogenesis), and KEGG category ko03036 (chromosome and associated proteins) were invoved, suggesting functional divergence in cell division and biogenesis. The convergence of the differentiation signal is further supported by a correlation of 0.41 between PC1 gene loadings in the two species and an overlap of the majority of the extreme PC1 loadings genes in the D, M and ko03036 categories (**Fig. 4c**).

To characterize these putative ecospecies, we constructed phylogenetic trees based on SNPs with extreme loadings. We found two putative ecospecies for *S. mitis* and three for *S. oralis*, which also correspond to distinct clusters in strain PCA loadings (**Fig. 4a,b**). To determine whether they had evolved independently, we constructed a joint tree for the shared high loading genes, which imply independent origin of ecospecies in the two species (**Fig. 4c**). Finally, we constructed phylogenetic trees based on remaining SNPs and all SNPs to examine the pattern of differentiation in the rest of the genome. Distinct ecospecies clusters are visible implying modest genome-wide differentiation, which contrasts with the absence of differentiation in most of the genome seen in *V. parahaemolyticus* and *H. pylori*[37,39]. This result is consistent with the *Streptococcus* ecospecies being older and futher along the path of functional differentiation and helps to explain the higher long-range LD seen in these two species.

Discussion

Exploring and understanding the diversity of life requires a framework akin to a geographical map. For prokaryotes, classical resources such as *Bergey's Manual* have long served as the initial "map". MAP extends this foundation with a genome-based atlas, capturing rich genomic and population genetic features across prokaryotic species. It is accessible online, enabling straightforward exploration and use. MAP serves both focused and broad research needs. For researchers interested in particular microbes, it provides concise summaries of genomic, population genetic, phenotypic, and ecological features. Beyond single species, it facilitates straightforward comparisons across all available prokaryotes and serves as a resource for macrogenomic and population genetic studies, including both raw and curated non-redundant datasets.

Our analysis of MAP has generated a number of interesting observations that deserve further study. For example,we find that prokaryotic cell size negatively correlated with π_s, which is in the same direction as a stronger negative correlation between propagule size and diversity in eukaryotes[10]. Bacterial genome size approximates a normal distribution, but unexpectedly, core- and pan-gene counts are bimodal. Gut-dominant *Bacteroides*, such as *B. uniformis* and *B. fragilis* have exceptionally high genome fluidity. Lastly, we confirm that selection favors bigger genomes, and also find a negative correlation between gene length and *dN/dS*, suggesting that selection also favors long genes.

Our analyses and resources are available based on two taxonomic classifications, namely NCBI and Genome Taxonomy Database (GTDB)[40,41]. The problem of bacterial species definitions remains unsolved and large databases inevitably contain errors[42,43]. NCBI is based on user inputted classifications, while GTDB uses an automatic algorithm, based on a universal average nucleotide identity cutoff of 0.95. While GTDB has advantages of consistency and quality control over NCBI, it often splits taxonomically diverse named species into multiple putative species (Supplementary Fig. 5). In many cases, these species do not correspond to coherent evolutionary units. In particular, highly diverse species such as *P. fluorescens* have very low short-range LD, indicating that gene flow is effective in distributing diversity within the species, and accordingly they are amongst the best examples of biological species, despite being split into >10 species by GTDB. Therefore GTDB is not especially well-suited for macrogenomic analysis, especially for highly diverse species, and we presented our main results based on NCBI classification. We nevertheless provide an alternative version of parameter calculation and correlation analysis results based on GTDB (Supplementary Tables 7-9), and show a strong correlation between parameter relationships derived from two taxonomies (r = 0.96, Supplementary Fig. 6), indicating that the great majority of our results are robust to species definitions.

We highlight the capacity of MAP to reveal novel observations, to promote the generation and testing of new hypotheses, and thereby to refine existing theoretical frameworks. We show that short- and long-range genetic linkage capture distinct aspects of evolutionary dynamics and contribute independently to genetic diversity, contradicting the predictions of neutral theory. Specifically, there is a pronounced positive correlation between long-range LD and π_s for species exhibiting similar levels of short-range LD. This correlation seems to underly a paradoxical feature, namely that despite there being highly diverse species in the data, panmictic species with low LD at all distances are absent, suggesting they are also rare or absent in nature[33]. Our resolution of the paradox is that as within-species diversity increases, epistatic selection becomes an increasingly strong force structuring diversity.

An important implication of our finding is that we can get insight into the ecological pressures acting on genetically diverse prokaryotic species, simply by characterizing how their diversity is structured. One example of structure generated by selection is bacterial ecospecies, which are characterized by fixed differences in a small fraction of the genome, but high recombination elsewhere. We have previously described the first two bacterial ecospecies in *V. parahaemolyticus* and *H. pylori*. In *V. parahaemolyticus* the structure reflects diversification at motility related genes, while in *H. pylori* differentiation has occurred in genes related to immune escape and nutrient

uptake [37,39]. Both species have moderate diversity and ecospecies structure is restricted to a small fraction of the genome, representing an early stage of diversification.

Here we have identified two additional species with ecospecies-like structure, namely *S. mitis* and *S. oralis*. These two sister taxa both inhabit the mouths of humans and other animals and there is a substantial convergence in the genetic basis of the ecospecies structure, with the same cell division and biogenesis associated genome categories and genes involved, suggesting similar ecological driving forces and/or adaptive potentials in both species, which might reflect adaptation to the highly dynamic oral environment. The ecospecies appear to be older than those identified previously and show evidence for some level of differentiation genome-wide, suggesting a later stage of diversification and is likely to be a precursor to full speciation.

Extrapolating from early to late stage ecospecies divergence and beyond, we can begin to reconstruct the life-cycle of many prokaryotic species. However, ecospecies is only one type of structure caused by selection. We have previously hypothesized the existence of bacterial ecoclines, which exhibit continuous variation driven by a quantitative trait[44] and the genetic structure in the most diverse species such as *P. fluorescens* resists any immediate classification. Characterizing the full range of prokaryotic genetic structures should provide a unique new window into the forces shaping the most diverse and adaptable organisms on our planet.

Methods

Genome datasets

Deduplication and quality control

We downloaded all assembled genomes (n=317,542) included in the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org, release 207)[40,41]. We performed deduplications based on the Biosample accessions, retaining only the genome with highest N50 for each Biosample. High-quality genomes with completeness <90% or contamination >5% (estimated by CheckM[45]) were excluded.

The remaining genomes (n=268,463) were grouped into species based on National Center for Biotechnology Information (NCBI) taxonomy and GTDB classifications. For each species, the earliest sequenced complete genomes or the genome with highest N50 (if no complete genome was available) were selected as the reference genomes. All genomes of a species were aligned against the reference using MUMmer v3.23[46] to generate whole-genome alignments. Genomes with <60% coverage against the reference (n=4,280) were excluded.

Representative datasets and clonal groups

While reliable estimation of population genetic parameters relies on random representative datasets, biased repeat sampling (*e.g.*, strains from disease outbreaks) is common for public datasets, especially for pathogenic species. To reduce the effect of repeat sampling, we generated representative genome datasets. For each species, core-genome (regions present in >95% strains) single-nucleotide polymorphisms (SNPs) were identified using SNP-sites v2.5.1[47]. SNPs located in repetitive regions identified by TRF v4.07b[48] or BLASTN self-search were excluded. Pairwise SNP-distance between strains were calculated using snp-dists v0.8.2

(https://github.com/tseemann/snp-dists). Genomes of a species were assigned to complete-linkage clusters (CLCs) based on SNP-distance. For representative datasets of each species, we used a "relative value" - 1% median pairwise SNP-distance as the cutoff, and selected one genome with highest N50 from a CLC or these singletons as representative genomes.

Substitution rate and part of recombination parameter estimations rely on clonal groups (CGs) consisting of closely related genomes. We used an "absolute value" – 5,000 SNPs (total SNP threshold per branch for reliable recombination detection[49]) as the cutoff to identify CLCs representing CGs.

Genomic parameters

We calculated 12 categories of genomic parameters (GP) for all representative genomes, using the median value of representative genomes to represent each species. The variability (standard deviation [SD]/median ratio) of each parameter is shown in Supplementary Fig. 2. Some GP estimates exhibited relatively high variability (SD/median ratio >0.5), which we interpreted as reflecting substantial within-species heterogeneity among strains. Rather than excluding these parameters, we retained them and used the median value as a representative measure, in order to maximize comparability across species.

Genome size and GC content were calculated using Quast v5.0.2[50]. We re-annotated representative genomes using Prokka v1.13[51] with default setting except for the "gcode" (genetic code) option. Genetic codes for each species were obtained from the NCBI taxonomy database. For species included in MAP, we found that all species use translation table 11 (Bacterial, Archaeal, and Plant Plastid Code), except for species classified under NCBI taxonomy "Mycoplasmatales and Entomoplasmatales" (153 species) or GTDB taxonomy "Mycoplasmatales" (198 species), which use translation table 4. Gene number and coding proportion were extracted or calculated based on annotation results. Codon bias parameters (CBI, Fop, ENc) were estimated using codonW v1.3 (https://codonw.sourceforge.net/) based on coding sequences. The number of tRNA and rRNA genes was extracted from Prokka annotations.

The standard protein sequences of restriction-modification (R-M) system genes were retrieved from REBASE[52] and searched against annotated protein sequences to identify R-M genes using BLASTp; hits with >50% coverage were considered present as previously described[36]. Repetitive regions of genomes were identified using TRF v4.07b[48] and BLASTN self-search. CRISPR arrays were identified using CRISPRidentify v1.1.0[53]. Phage sequences were identified using PhiSpy v3.7.8[54]. Insertion sequence (IS) elements were identified using ISEScan v1.7.2.3[55]. Default parameters were used for these analysis unless otherwise specified.

Population genetic parameters

We calculated seven categories of population genetic parameters (PGP) for species with ≥10 representative genomes and for CGs with ≥10 genomes.

Random sampling procedure

To reduce potential biases arising from unequal numbers of representative genomes, five categories of parameters (π_s, *dN/dS*, LD r^2, *h/m*, core-/pan-genome and gene fluidity) were estimated using random sampling of 10 representative genomes. Sampling was repeated 10 times independently.

For *dN/dS* and LD r^2, which require specific thresholds for reliable estimation (see below), some replicates failed to yield results. Only estimates with ≥5 successful replicates were retained for downstream analyses. To further ensure robustness, we excluded cases with unstable outcomes, defined as those with a SD/median ratio >0.5 across replicates. The final estimate for each species was taken as the median across retained replicates. The variability (SD/median ratio) of each parameter is shown in Supplementary Fig. 2.

π_s and dN/dS

For each species, coding sequence alignments of representative genomes were extracted from whole-genome alignment. Pairwise non-synonymous divergence (*dN*), synonymous divergence (*dS*), and *dN/dS* were calculated with PAML v4.10.0[56] using the yn00 algorithm. The mean pairwise *dS* was used to represent the synonymous nucleotide diversity (π_s). For *dN/dS* estimation, to reduce the impact of heterogeneous divergence levels across species while maximizing the number of species that could be compared, we retained only genome pairs with *dS* values within 0.03–0.13, corresponding to the interquartile range (IQR) across all species. For random sampling analyses, only *dN/dS* estimates based on >10 strain pairs were retained.

Linkage disequilibrium

Linkage disequilibrium (LD) correlation coefficient between two loci (r^2) at distances of 1 bp, 10 bp, 100 bp, 1 kb, and 10 kb was calculated using Haploview v4.2[57]. For each replication, estimates were retained only when >1,000 locus pairs were available at the given distance. We also calculated the LD satuation level, defined as the ratio of r^2 between different distances (r^2 at 10 kb / r^2 at 10 bp).

h/m

We calculated the ratio of homoplasic and non-homoplasic alleles (*h/m*) using ConSpeciFix v1.3.0[58] based on randomly sampled and all representative genomes.

Core- and pan-genome and gene fluidity

For each species, the number of core- and pan-genes was calculated using Panaroo v1.2.8[59] based on re-annotated results. Gene fluidity was measured by the mean pairwise ratio between the number of genes not shared and the total number of genes.

Recombination estimation based on all representative genomes

Five recombination-related parameters were estimated from all representative genomes using mcorr[16] with default settings. These included mutational divergence (θ_{pool}), recombinational divergence (φ_{pool}), mean fragment size (δ), fraction of the sample diversity resulting from recombination (c), and the relative effect of recombination to mutation ($r/m = \varphi_{\text{pool}}/\theta_{\text{pool}}$).

Recmobination level and substitution rate of clonal group (CG)

For each CG, all available genomes were used to estimate recombination-related parameters and substitution rates. The variability (SD/median ratio) of each parameter across CGs within a species is shown in Supplementary Fig. 2. We interpreted high variability (SD/median ratio >0.5) as reflecting substantial within-species heterogeneity among CGs. Rather than discarding such cases, we retained them and used the median value as the representative estimate for each species, thereby maximizing comparability across species.

Four recombination-related parameters were estimated for each CG using ClonalFrameML v1.12[28]: the relative rate of recombination to mutation (R/θ), mean divergence between donor and recipient (v), δ and r/m. In addition, h/m was estimated using ConSpeciFix v1.3.0[58].

For each CG of a species, we reconstructed maximum-likelihood (ML) phylogenetic trees using FastTree v 2.1.10[60] based on non-repetitive and non-recombinational SNPs as previously described[61]. The isolation year of strains were extracted from the corresponding BioSample information. We used the ML tree and isolation year information to calculate the mutation rate with BactDating v1.1.0[62] as previously described[63]. Only substitution rates with effective sample sizes (ESS) >100 were retained.

Phenotype and ecology data

The phenotype and ecology data were extracted from three sources: Madin *et al*[64], Secaira-Morocho *et al*[65] and BacDive database (up to August 2025)[66]. Data from the first two sources were available at the species level. For species with multiple strain entries, quantitative traits were summarized using the mean or median, while categorical traits were assigned according to the majority state (term occurring in >50% of strains).

BacDive provides strain-level data, which were aggregated to the species level using the same approach: the median for quantitative traits and the majority state for categorical traits. When no single state exceeded 50% frequency, the categorical trait was recorded as “not available (NA)”. For BacDive-derived traits, within-species variability (SD/median ratio across strains) for quantitative traits and the frequencies of majority states across species for categorical traits are shown in Supplementary Fig. 2. Traits with high variability were retained to maximize comparability across species.

We further combined data from all three sources by taking the median of quantitative traits and the majority state across sources for categorical traits. The resulting species-level data from each individual source, as well as the combined dataset, are presented in Supplementary Table 3.

Correlation analysis

We used the Python packages pingouin and scipy.stats to calculate correlations between different parameter types. Specifically, we generated all pairwise combinations of parameters and performed correlation analyses on parameter pairs containing more than 100 observations. For quantitative data, we applied three methods to compute the correlation coefficient (r): Pearson, Spearman, and Kendall. For categorical data, we performed logistic regression and quantified

associations with categorical variables using the eta coefficient (η). When significant correlations were detected ($p < 0.01$ and coefficient r or $\eta > 0.1$), the highest correlation coefficient was retained to represent the relationship between the parameters.

We conducted phylogenetic signal analysis on quantitative data parameters using phylolm v2.6.2[67]. The phylogenetic trees based on universally distributed genes (archaea: 53 genes, bacteria: 120 genes) used by GTDB[40,41] were employed in the analysis. Pagel's λ value was used to quantify the strength of the phylogenetic signal.

Principal component analysis (PCA)

PCA was performed using PLINK v1.9[68] based on core-SNPs with a minor allele frequency (MAF) threshold of 2%. The SNP loading for each principal component (PC) was calculated as described in Liu *et al.*[44], reflecting the correlation between SNPs and PCs. Briefly, strain loadings were derived by multiplying the eigenvectors by the square root of the eigenvalues corresponding to each PC. SNP loadings were then computed by matrix multiplication between strain loadings and the SNP matrix.

SNP annotation and gene enrichment analysis

Synonymous and non-synonymous SNPs were annotated using SnpEff v4.3t[69], based on the re-annotated reference genome. Genes associated with SNPs in the top 1% of SNP loadings on PC1 were extracted for subsequent analysis. Functional enrichment was performed using the R package ClusterProfiler v4.2.2[70].

Between-species homologous gene analysis

Homologous genes between *Streptococcus mitis* and *S. oralis* were identified using BLASTp based on annotated protein sequences, with a minimum identity and coverage of 50%. Protein sequences of homologous genes associated with SNPs in the top 1% of SNP loadings on PC1 were used for phylogenetic tree construction using FastTree v2.1.10[60].

Data availability

All genome sequences are available from the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org, release 207). Essential data files for replication, including whole-genome alignments, the SNP matrix, sorted metadata, and calculated results, are available at www.genomap.cn, Zenodo (https://zenodo.org/records/14880463), and provided in Supplementary Tables. Source data are provided with this paper.

References

1. Hug, L. A. *et al*. A new view of the tree of life. *Nat Microbiol* **1**, 1–6 (2016).
2. Rainey F. *et al*. Bergey's manual of systematics of archaea and bacteria. Hoboken, NJ: *Wiley*, (2015). doi:10.1002/9781118960608.
3. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17** (2016).
4. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nature Reviews Genetics* **22** (2021).

5. Blanchet, S., Prunier, J. G. & De Kort, H. Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends in Genetics* **33**, 579–580 (2017).
6. Leigh, D. M. *et al.* Opportunities and challenges of macrogenetic studies. *Nature Reviews Genetics* **22**, 791–807 (2021).
7. Manel, S. *et al.* Global determinants of freshwater and marine fish genetic diversity. *Nat Commun* **11** (2020).
8. Theodoridis, S. *et al.* Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nat Commun* **11** (2020).
9. Vellend, M. & Geber, M. A. Connections between species diversity and genetic diversity. *Ecology Letters* **8** (2005).
10. Romiguier, J. *et al.* Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261–263 (2014).
11. Miraldo, A. *et al.* An Anthropocene map of genetic diversity. *Science (1979)* **353**, (2016).
12. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc Natl Acad Sci U S A* **115**, 6506–6511 (2018).
13. Denamur, E., Clermont, O., Bonacorsi, S. & Gordon, D. The population genetics of pathogenic Escherichia coli. *Nat Rev Microbiol* **19**, 37–54 (2021).
14. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10** (2009).
15. Suerbaum, S. *et al.* Free recombination within Helicobacter pylori. *Proc Natl Acad Sci U S A* **95**, 12619-12624 (1998).
16. Lin, M. & Kussell, E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods* **16**, 199–204 (2019).
17. Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc Natl Acad Sci U S A* **113**, 11399–11407 (2016).
18. Lynch, M. & Conery, J. S. The Origins of Genome Complexity. *Science (1979)* **302**, 1401–1404 (2003).
19. Agnoli, K. *et al.* Exposing the third chromosome of Burkholderia cepacia complex strains as a virulence plasmid. *Mol Microbiol* **83** (2012).
20. Lin, L. F., Posfai, J., Roberts, R. J. & Kong, H. Comparative genomics of the restriction-modification systems in Helicobacter pylori. *Proc Natl Acad Sci U S A* **98** (2001).
21. Weigand, M. R. *et al.* The history of Bordetella pertussis genome evolution includes structural rearrangement. *J Bacteriol* **199** (2017).
22. Zafar, H. & Saier, M. H. Gut Bacteroides species in health and disease. *Gut Microbes* **13** (2021).
23. Celli, J. The Intracellular Life Cycle of Brucella spp . *Microbiol Spectr* **7** (2019).
24. Teng, W., Liao, B., Chen, M. & Shu, W. Genomic Legacies of Ancient Adaptation Illuminate GC-Content Evolution in Bacteria. *Microbiol Spectr* **11** (2023).
25. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Molecular Biology Reports* **49** (2022).
26. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol* **5**, 966–977 (2013).

27. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nature Reviews Genetics* **17**, 422–433 (2016).
28. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput Biol* **11**, 1–18 (2015).
29. Fuhrman, J. A. *et al.* A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci U S A* **105** (2008).
30. Andam, C. P. *et al.* A latitudinal diversity gradient in terrestrial bacteria of the genus Streptomyces. *mBio* **7** (2016).
31. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* **6** (2010).
32. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* **113**, 5658–5663 (2016).
33. Yang, C., Cui, Y., Didelot, X., Yang, R. & Falush, D. Why panmictic bacteria are rare. *bioRxiv* (2018) doi:10.1101/385336.
34. Didelot, X., Méric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics* **13** (2012).
35. Thorpe, H. A. *et al.* Repeated out-of-Africa expansions of Helicobacter pylori driven by replacement of deleterious mutations. *Nat Commun* **13** (2022).
36. Yang, C. *et al.* Recent mixing of Vibrio parahaemolyticus populations. *ISME J* **13**, 2578–2588 (2019).
37. Cui, Y. *et al.* The landscape of coadaptation in Vibrio parahaemolyticus. *Elife* **9**, e54136 (2020).
38. Svensson, S. L. *et al.* Evolution of hunt, kill, devour in a Vibrio ecospecies. *bioRxiv* (2025) doi:10.1101/2025.06.27.661923.
39. Tourrette, E. *et al.* An ancient ecospecies of Helicobacter pylori. *Nature* **635**, 178–185 (2024).
40. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996 (2018).
41. Parks, D. H. *et al.* GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785–D794 (2022).
42. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiol Rev* **25** (2001).
43. VanInsberghe, D., Arevalo, P., Chien, D. & Polz, M. F. How can microbial population genomics inform community ecology? *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190253 (2020).
44. Liu, S., Svensson, S. L. & Falush, D. A putative ecocline in Klebsiella pneumoniae. *bioRxiv* (2025) doi:10.1101/2025.01.20.633859.
45. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043–1055 (2015).
46. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to Identify Similar Regions in Large Sequence Sets . *Curr Protoc Bioinformatics* **1**, 10-3 (2003).

47. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016).
48. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
49. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).
50. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
51. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
52. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **51** (2023).
53. Mitrofanov, A. *et al.* CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res* **49** (2021).
54. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res* **40** (2012).
55. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33** (2017).
56. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24** (2007).
57. Barrett, J. C. Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc* **4** (2009).
58. Bobay, L. M., Ellis, B. S. H. & Ochman, H. ConSpeciFix: Classifying prokaryotic species based on gene flow. *Bioinformatics* **34** (2018).
59. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
60. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).
61. Yang, C. *et al.* Outbreak dynamics of foodborne pathogen Vibrio parahaemolyticus over a seventeen year period implies hidden reservoirs. *Nat Microbiol* **7**, 1221–1229 (2022).
62. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res* **46**, (2018).
63. Yang, C. *et al.* Wave succession in the pandemic clone of Vibrio parahaemolyticus driven by gene loss. *Nat Ecol Evol* (2025) doi:10.1038/s41559-025-02827-z.
64. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* **7** (2020).
65. Secaira-Morocho, H., Chede, A., Gonzalez-de-Salceda, L., Garcia-Pichel, F. & Zhu, Q. An evolutionary optimum amid moderate heritability in prokaryotic cell size. *Cell Rep* **43**, 114268 (2024).
66. Reimer, L. C. *et al.* BacDive in 2022: The knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res* **50** (2022).
67. Tung Ho, L. S. & Ané, C. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst Biol* **63** (2014).

68. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
69. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6** (2012).
70. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

Acknowledgements

This work was supported by National Key Research and Development Program of China (No. 2022YFC2304700 to F.D.), National Natural Science Foundation of China (No. 32270003 and No. 32000008 to C.Y., No. 32170640 and No.32211550014 to D.F.), Youth Innovation Promotion Association, Chinese Academy of Sciences (No. 2022278 to C.Y.) and Shanghai Rising-Star Program (No. 23QA1410500 to C.Y.).

Author contributions

C.Y., D.F., R.Y., and Y.C. initiated and coordinated the study. C.Y., N.W., H.H. and W.X. contributed to data collectionand and performed bioinformatics analysis. F.J. and S.T. contributed to website construction and visualization. All authors contributed to interpretation of the data. C.Y. and D.F. wrote the first draft of the paper and D.X., R.Y., and Y.C. reviewed and revised the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

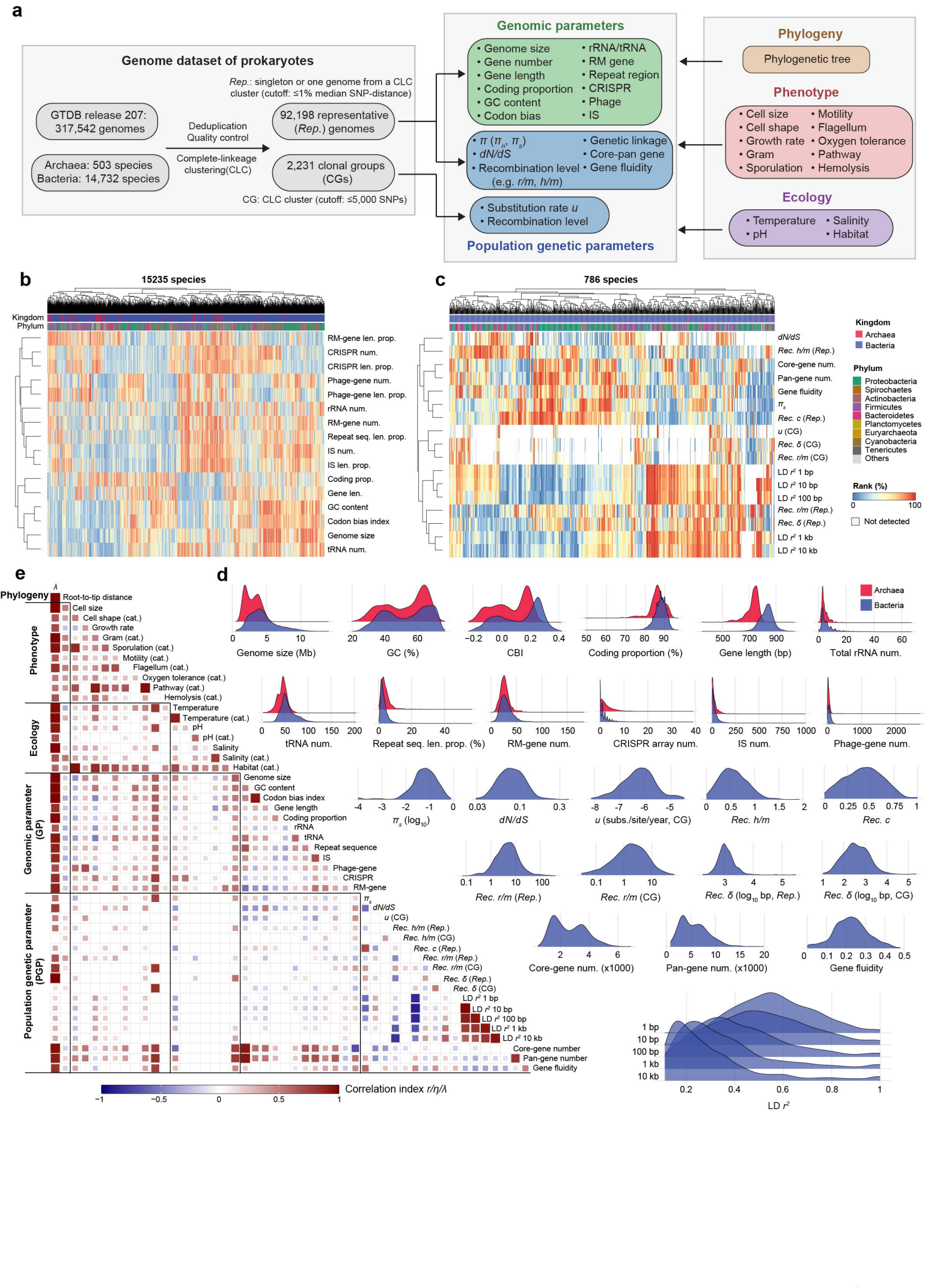
Figure legends

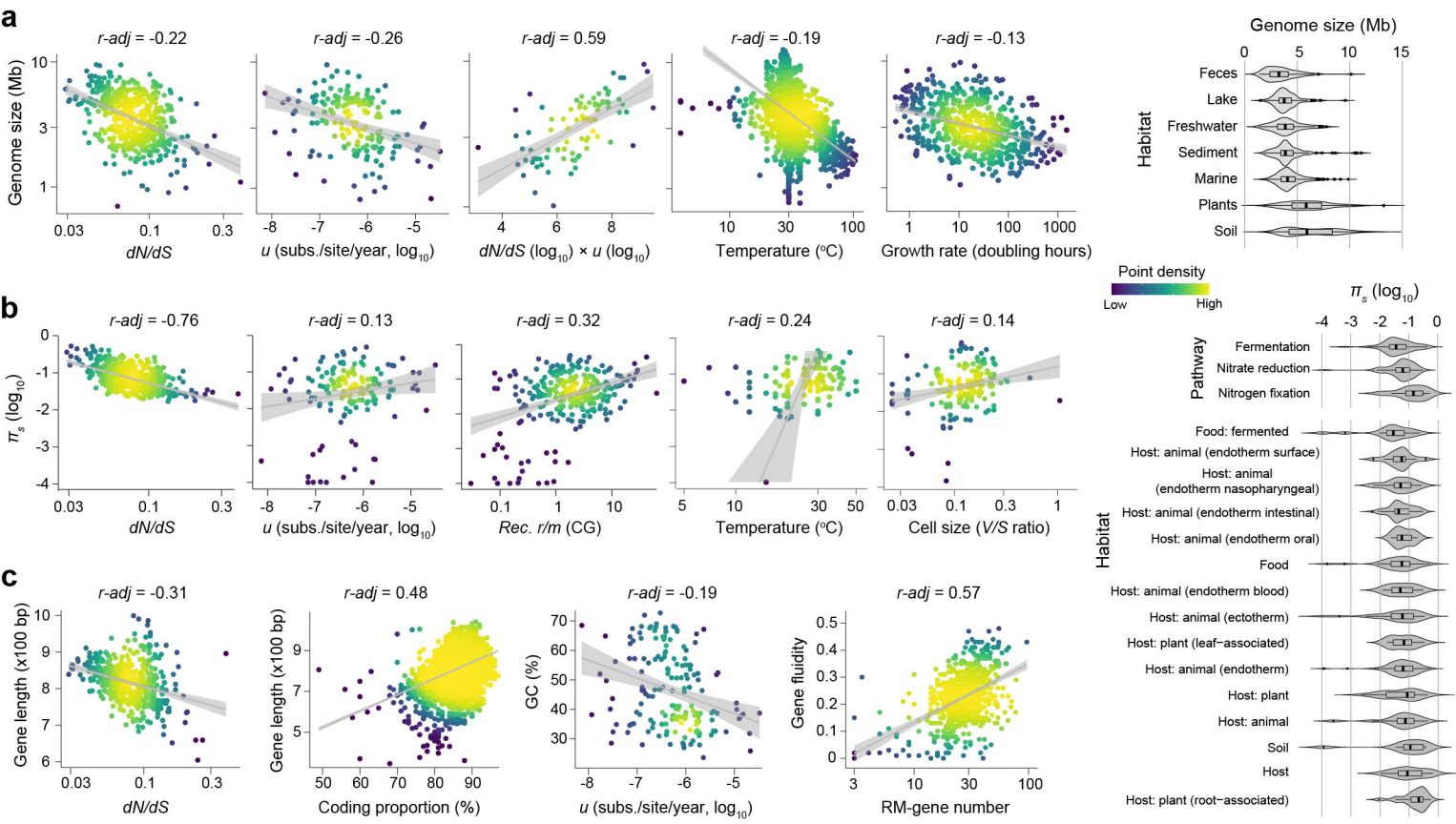
Figure 1. Overview of the MAP. **a)** Study flowchart. **b,c)** Heatmaps showing the rank percentages of genomic parameters (GP) and population genetic parameters (PGP) for each species. Bars at the top indicate the kingdom and phylum, while heatmap colors represent rank percentages. **d)** Distribution of different GP and PGP. Red and blue represent archaea and bacteria, respectively. PGP is shown only for bacteria, as few archaeal species met the criterion for PGP estimation (≥10 representative genomes). **e)** Correlation heatmap among various parameters. Phylogenetic signals were quantified using Pagel's λ. *Rep.*: representative dataset. CG: clonal group dataset. num.: number. len.: length, prop.: proportion. *Rec.*: recombination-ralated parameters. *c*: recombination coverage. δ: recombination fragment size. *r/m*: relative effect of recombination to mutation. *u*: substitution rate. *h/m*: ratio of homoplasic and non-homoplasic alleles. *dN/dS*: ratio of non-synonymous divergence (*dN*) to synonymous divergence (*dS*). π_s: synonymous genetic diversity. cat.: categorical data.

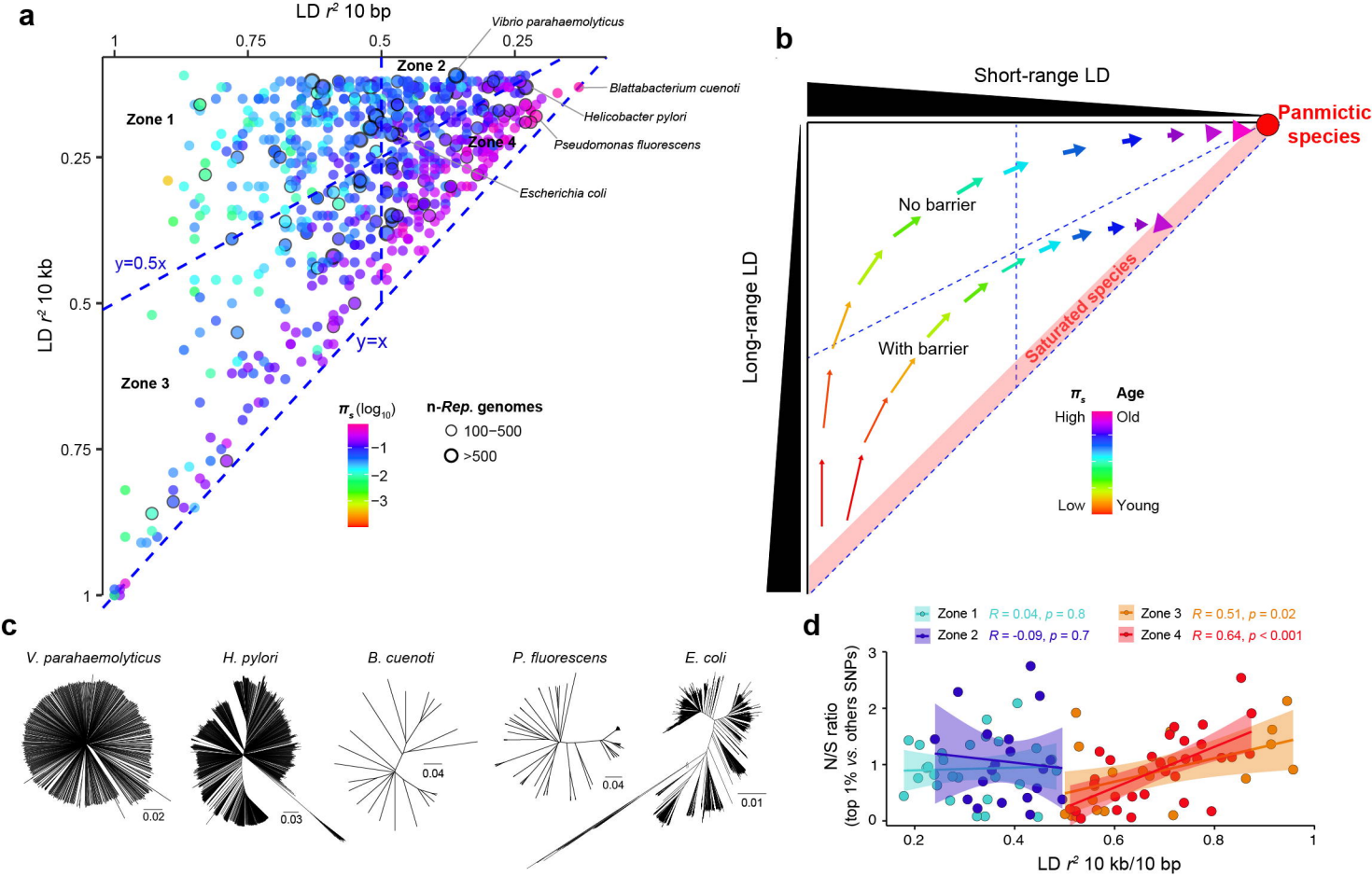
Figure 2. Correlations related to genome size and genetic diversity, and underexplored associations. a) Correlations associated with genome size. **b)** Correlations associated with genetic diversity. c) Underexplored correlations revealed by MAP. *r-adj* represents the correlation coefficient after adjustment for phylogenetic signal. Point colors represent point density. The grey line shows the fitted linear relationship, and the shaded area represents the 95% confidence interval. For qualitative variables, only categories with ≥10 species are shown.

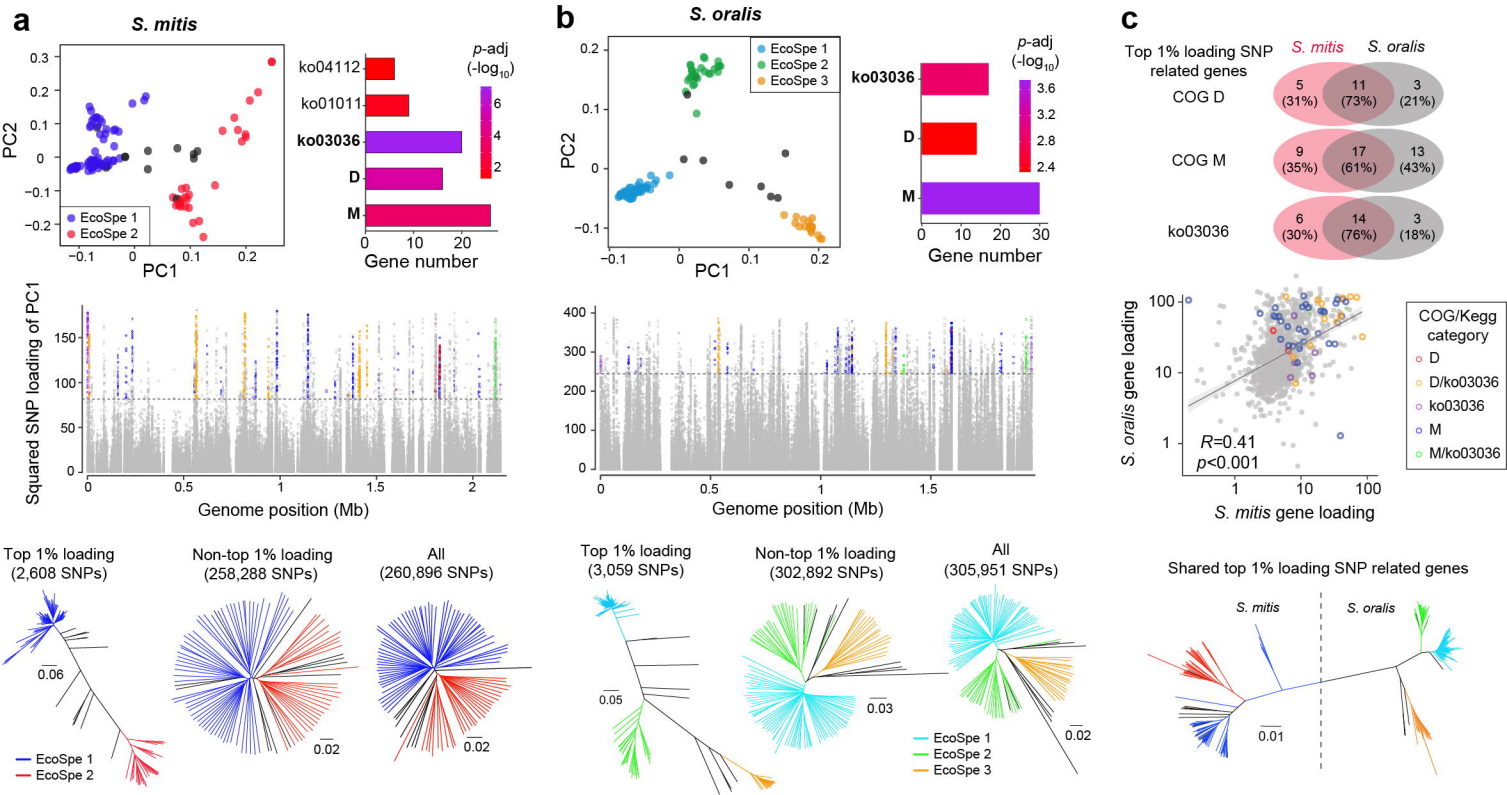
Figure 3. Overview of prokaryotic genetic diversity. a) Relationships between short- (10 bp), long-range LD (10 kb), and genetic diversity. Point colors represent genetic diversity levels. Species with ≥100 representative genomes are outlined in black. Blue dashed lines indicate LD saturation levels of 0.5 (y=0.5x) and 1 (y=x). **b)** Conceptual model illustrating the evolution of LD, genetic diversity, and population structure. The x- and y-axes represent the strengths of short- and long-range LD, respectively. Colored arrows depict evolutionary trajectories of species with or without genetic barriers, with colors indicating species ages reflected in genetic diversity. Saturated species indicate those with similar levels of short- and long-range LD. Panmictic species denote freely recombining, saturated, and highly diverse species. **c)** Phylogenetic trees of five representative species. **d)** Correlations between LD saturation level and N/S ratio (top 1% *vs*. other SNPs), representing the ratio of non-synonymous to synonymous sites (N/S) for SNPs in the top 1% of SNP loadings on PC1 (SNP-top1%-PC1) relative to that for SNPs outside the top 1%.

Figure 4. Covergent ecospecies. a,b) Top: Principal component analysis (left; point colors indicate inferred ecospecies) and functional enrichment of genes associated with SNP-top1%-PC1 (right; as defined in Fig. 3d) in *Streptococcus mitis* and *S. oralis*. Middle: Distribution of squared SNP loadings on PC1 across the genome; point colors represent COG or KEGG categories. Dashed lines indicate the top 1% SNP loading cutoff. Bottom: Phylogenetic trees based on SNPs within the top 1% SNP loadings on PC1, remaining and all SNPs. Branch colors indicate inferred ecospecies. **c)** Top: Shared homologous genes associated with SNP-top1%-PC1 between the two species. Middle: Correlation of gene loading (mean SNP loading per gene) between species; point colors represent COG or KEGG categories. Bottom: Phylogenetic trees based on protein sequences of shared genes associated with SNP-top1%-PC1 between species.


a
Genome dataset of prokaryotes
Rep.: singleton or one genome from a CLC cluster (cutoff: ≤1% median SNP-distance)
GTDB release 207: 317,542 genomes
Archaea: 503 species
Bacteria: 14,732 species
Deduplication
Quality control
Complete-linkeage clustering(CLC)
92,198 representative (Rep.) genomes
2,231 clonal groups (CGs)
CG: CLC cluster (cutoff: ≤5,000 SNPs)
Genomic parameters
• Genome size • rRNA/tRNA
• Gene number • RM gene
• Gene length • Repeat region
• Coding proportion • CRISPR
• GC content • Phage
• Codon bias • IS
• π (πn, πs) • Genetic linkage
• dN/dS • Core-pan gene
• Recombination level (e.g. r/m, h/m) • Gene fluidity
• Substitution rate u
• Recombination level
Population genetic parameters
Phylogeny
Phylogenetic tree
Phenotype
• Cell size • Motility
• Cell shape • Flagellum
• Growth rate • Oxygen tolerance
• Gram • Pathway
• Sporulation • Hemolysis
Ecology
• Temperature • Salinity
• pH • Habitat
b
15235 species
Kingdom
Phylum
RM-gene len. prop.
CRISPR num.
CRISPR len. prop.
Phage-gene num.
Phage-gene len. prop.
rRNA num.
RM-gene num.
Repeat seq. len. prop.
IS num.
IS len. prop.
Coding prop.
Gene len.
GC content
Codon bias index
Genome size
tRNA num.
c
786 species
dN/dS
Rec. h/m (Rep.)
Core-gene num.
Pan-gene num.
Gene fluidity
πs
Rec. c (Rep.)
u (CG)
Rec. δ (CG)
Rec. r/m (CG)
LD r2 1 bp
LD r2 10 bp
LD r2 100 bp
Rec. r/m (Rep.)
Rec. δ (Rep.)
LD r2 1 kb
LD r2 10 kb
Kingdom
Archaea
Bacteria
Phylum
Proteobacteria
Spirochaetes
Actinobacteria
Firmicutes
Bacteroidetes
Planctomycetes
Euryarchaeota
Cyanobacteria
Tenericutes
Others
Rank (%)
0
100
Not detected
e
Phylogeny
λ
Root-to-tip distance
Phenotype
Cell size
Cell shape (cat.)
Growth rate
Gram (cat.)
Sporulation (cat.)
Motility (cat.)
Flagellum (cat.)
Oxygen tolerance (cat.)
Pathway (cat.)
Hemolysis (cat.)
Ecology
Temperature
Temperature (cat.)
pH
pH (cat.)
Salinity
Salinity (cat.)
Habitat (cat.)
Genomic parameter (GP)
Genome size
GC content
Codon bias index
Gene length
Coding proportion
rRNA
tRNA
Repeat sequence
IS
Phage-gene
CRISPR
RM-gene
Population genetic parameter (PGP)
πs
dN/dS
u (CG)
Rec. h/m (Rep.)
Rec. h/m (CG)
Rec. c (Rep.)
Rec. r/m (Rep.)
Rec. r/m (CG)
Rec. δ (Rep.)
Rec. δ (CG)
LD r2 1 bp
LD r2 10 bp
LD r2 100 bp
LD r2 1 kb
LD r2 10 kb
Core-gene number
Pan-gene number
Gene fluidity
−1
−0.5
0
0.5
1
Correlation index r/η/λ
d
Archaea
Bacteria
Genome size (Mb)
GC (%)
CBI
Coding proportion (%)
Gene length (bp)
Total rRNA num.
tRNA num.
Repeat seq. len. prop. (%)
RM-gene num.
CRISPR array num.
IS num.
Phage-gene num.
πs (log10)
dN/dS
u (subs./site/year, CG)
Rec. h/m
Rec. c
Rec. r/m (Rep.)
Rec. r/m (CG)
Rec. δ (log10 bp, Rep.)
Rec. δ (log10 bp, CG)
Core-gene num. (x1000)
Pan-gene num. (x1000)
Gene fluidity
1 bp
10 bp
100 bp
1 kb
10 kb
LD r2


a
r-adj = -0.22
r-adj = -0.26
r-adj = 0.59
r-adj = -0.19
r-adj = -0.13
Genome size (Mb)
dN/dS
u (subs./site/year, log10)
dN/dS (log10) × u (log10)
Temperature (°C)
Growth rate (doubling hours)
Genome size (Mb)
Habitat
Feces
Lake
Freshwater
Sediment
Marine
Plants
Soil
Point density
Low
High
b
r-adj = -0.76
r-adj = 0.13
r-adj = 0.32
r-adj = 0.24
r-adj = 0.14
πs (log10)
dN/dS
u (subs./site/year, log10)
Rec. r/m (CG)
Temperature (°C)
Cell size (V/S ratio)
πs (log10)
Pathway
Fermentation
Nitrate reduction
Nitrogen fixation
Habitat
Food: fermented
Host: animal (endotherm surface)
Host: animal (endotherm nasopharyngeal)
Host: animal (endotherm intestinal)
Host: animal (endotherm oral)
Food
Host: animal (endotherm blood)
Host: animal (ectotherm)
Host: plant (leaf-associated)
Host: animal (endotherm)
Host: plant
Host: animal
Soil
Host
Host: plant (root-associated)
c
r-adj = -0.31
r-adj = 0.48
r-adj = -0.19
r-adj = 0.57
Gene length (x100 bp)
dN/dS
Gene length (x100 bp)
Coding proportion (%)
GC (%)
u (subs./site/year, log10)
Gene fluidity
RM-gene number


a
LD r² 10 bp
1
0.75
0.5
0.25
Zone 1
Zone 2
Zone 3
Zone 4
Vibrio parahaemolyticus
Blattabacterium cuenoti
Helicobacter pylori
Pseudomonas fluorescens
Escherichia coli
y=0.5x
y=x
LD r² 10 kb
0.25
0.5
0.75
1
πs (log10)
−1
−2
−3
n-Rep. genomes
100−500
>500
b
Short-range LD
Panmictic species
No barrier
With barrier
Saturated species
Long-range LD
πs
High
Low
Age
Old
Young
c
V. parahaemolyticus
H. pylori
B. cuenoti
P. fluorescens
E. coli
0.02
0.03
0.04
0.04
0.01
d
Zone 1 R = 0.04, p = 0.8
Zone 2 R = -0.09, p = 0.7
Zone 3 R = 0.51, p = 0.02
Zone 4 R = 0.64, p < 0.001
N/S ratio
(top 1% vs. others SNPs)
0
1
2
3
0.2
0.4
0.6
0.8
1
LD r² 10 kb/10 bp


a
S. mitis
PC2
PC1
EcoSpe 1
EcoSpe 2
ko04112
ko01011
ko03036
D
M
p-adj (-log10)
Gene number
Squared SNP loading of PC1
Genome position (Mb)
b
S. oralis
EcoSpe 1
EcoSpe 2
EcoSpe 3
c
Top 1% loading SNP related genes
S. mitis
S. oralis
COG D
5 (31%)
11 (73%)
3 (21%)
COG M
9 (35%)
17 (61%)
13 (43%)
ko03036
6 (30%)
14 (76%)
3 (18%)
S. oralis gene loading
S. mitis gene loading
R=0.41
p<0.001
COG/Kegg category
D
D/ko03036
ko03036
M
M/ko03036
Top 1% loading (2,608 SNPs)
Non-top 1% loading (258,288 SNPs)
All (260,896 SNPs)
Top 1% loading (3,059 SNPs)
Non-top 1% loading (302,892 SNPs)
All (305,951 SNPs)
Shared top 1% loading SNP related genes