Macrogenetic atlas of prokaryotes reveals selection-driven structures

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

[1] Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China;
[2] School of Public Health, Shanghai Jiao Tong University School of Medicine, 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.: yangchao@shsmu.edu.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. Further, 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. Since Linnaeus, taxonomy has provided this foundation based on phenotypic classification, and is now being systematically updated by the addition of genomic data, for example in the Tree of Life project[1]. The integration of genotype, phenotype, and ecological data remains the major unresolved challenge and is essential for characterizing species and identifying the forces driving diversification and speciation. Within microbiology, classical resources such as *Bergey's Manual*, which integrates taxonomy with phenotypic and ecological information[2], have long served this purpose.

Incorporating within-species genomic variation 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 are 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 are 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 seven 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 characterization 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 quantitative 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 range of prokaryotic diversity

Second, MAP delineates the range of genomic parameters across prokaryotes (**Fig. 1b-d**). For example, prokaryotes 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], define the upper 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

Fourth, 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**). We next show that the uniformly derived correlation matrices constitute an empirical resource for testing and refining evolutionary hypotheses.

Determinants of genetic diversity

Our understanding about the determinants of genetic diversity remains limited. N_e, mutation, and recombination have been proposed as major drivers[26]. 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 proxies such as the recombination-to-mutation ratio (r/m[27], *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. 2a**).

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[28,29] and eukaryotes[11].

Underexplored correlations

MAP also revealed previously overlooked patterns (**Fig. 2b**). Consistent with the established role of selection in shaping prokaryotic genome size[17], supported by its negative correlation with *dN/dS* (*r-adj* = -0.22), a proxy for selection efficacy, gene length shows an even stronger negative correlation with *dN/dS* (*r-adj* = -0.31). This pattern 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.

Moreover, substitution rate *u* is negatively correlated with GC content (*r-adj* = -0.19), consistent with a universal AT-biased mutation[30], and increases 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 their role in regulating DNA transfer[31].

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 higher 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. 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 (with non-recombining clonal species representing a complete barrier), reflected in long-range LD (**Fig. 3b**). As species accumulate age (indicated by increased π_s) in the absence of barriers, long-range LD decays rapidly, whereas short-range LD equilibrates more gradually, ultimately toward panmixia with high diversity and little LD at any distance (**Fig. 3b**, top arrows). By contrast, barriers to recombination slow the decay of long-range LD, raising its equilibrium level (**Fig. 3b**, middle and bottom arrows). Because LD decays more slowly in the presence of gene flow barriers, species positioned along a given vertical line in **Fig. 3a** tend to be "older" for higher long-range LD values (**Fig. 3b**), thereby explaining the positive association with π_s along each vertical line.

Although simplified, for example because species-wide bottlenecks can 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. For example, 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, and if maintained over longer timescales, diversity would increase considerably, driving the species toward panmixia[32].

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[33], 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 rates of

spread between human hosts as they dispersed out of Africa[34], 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**). 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.

Indeed, the species with the lowest short-range LD, namely *Blattabacterium cuenoti* (**Fig. 3a**), is not panmictic. As an ancient intracellular cockroach endosymbiont, it reproduces strictly asexually and is therefore expected to lack recombination[35]. Consistent with this, we are unable to find unambiguous evidence of recombination, as reflected by its "flat" LD decay curve (Supplementary Fig. 5). Instead, *B. cuenoti* appears mutation-saturated, exhibiting an extreme excess of triallelic and tetra-allelic SNP sites compared to other species (**Fig. 3d**), indicative of mutation saturation in an old and diverse species. Thus, its low LD arises from repeated mutation rather than recombination.

Epistatic selection as a driver of population structure

The paradoxical absence of true panmictic species can be explained if species with low barriers to exchange systematically evolve stronger barriers as species age and accumulate genetic diversity. *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. This predicts stronger signatures of selection in the genetic regions responsible for population structure in "older" species with higher genetic diversity. To test this, we performed principal component analysis (PCA) for each species with >100 genomes and identified SNPs most strongly associated with the first principal component (top 1% by loading values; PC1-related SNPs), representing the genetic basis of major population structure[39]. As a proxy for selection, we quantified the enrichment of non-synonymous sites among these SNPs relative to the genomic background, using the N/S (non-synonymous to synonymous site) ratio.

As predicted by the hypothesis, enrichment of non-synonymous sites was positively correlated with π_s, indicating that selection is most evident in highly diverse species (**Fig. 3e**). For example, *Pseudomonas fluorescens*, which exhibits very high genetic diversity, showed strong enrichment of non-synonymous sites (N/S ratio = 2.5) and a highly structured phylogeny (**Fig. 3c**). Together, these results support a model in which epistatic selection progressively shapes population structure

as species age and accumulate genetic diversity, thereby 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 genomes). Two such species are *Streptococcus mitis* and *S. oralis*. The PCA loading distribution for both resembled that seen for previous ecospecies[37,40], 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 involved, 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, together with substantial overlap among genes with extreme PC1 loadings 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 implies 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*[37,40]. This result is consistent with the *Streptococcus* ecospecies being older and further 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. The current MAP is primarily based on genomes from isolated strains, and future versions will incorporate metagenome-assembled genomes to expand species coverage and genomic diversity.

Our analysis of MAP has generated a number of interesting observations that deserve further study. For example, while bacterial genome size approximates a normal distribution, core- and pan-gene counts are unexpectedly bimodal. Gut-dominant *Bacteroides*, such as *B. uniformis* and *B. fragilis*, have exceptionally high genome fluidity. Moreover, 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)[41,42]. The problem of bacterial species definitions remains unsolved and large databases inevitably contain errors[43,44]. 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. 6). 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 among 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. 7), indicating that the great majority of our results are robust to species definitions.

The correlation matrix integrating genomic, population genetic, phenotypic, and ecological parameters provides a resource-level overview that facilitates hypothesis generation and testing, thereby helping to refine existing theoretical frameworks. These correlations, however, do not necessarily imply causal relationships and may include false positives arising from multiple testing or sensitivity to outliers; they should therefore be interpreted with caution. Nonetheless, they serve as a valuable starting point for hypothesis-driven and statistically rigorous follow-up analyses.

As an illustration, motivated by the observed lack of correlation between short- and long-range genetic linkage, we propose and provide evidence that these two linkage scales capture distinct aspects of evolutionary dynamics and independently shape genetic diversity. Classical population genetic theory predicts a negative correlation between genetic diversity and genetic linkage level[26]. We indeed find that short-range LD is strongly inversely correlated with genetic diversity. However, there is a pronounced positive correlation in MAP between long-range LD and π_s for species exhibiting similar levels of short-range LD (**Fig. 3a**). This arises because long-range LD principally reflects barriers to recombination, which leads to slower LD decay at all genetic distances (**Fig. 3b**). Therefore, for species with similar short-range LD values, species with higher long-range LD tend to be older and more diverse.

There is one zone in the triangle which is surprisingly empty, corresponding to panmictic species with low LD at all distances (**Fig. 3a**). Many bacteria occupy niches favorable to global dispersal and the maintenance of vast, stable populations over evolutionary timescales and this should generate panmictic, i.e., freely mixed species[32]. Our resolution of this paradox is that as within-species diversity increases, epistatic selection becomes an increasingly strong force. Consistent with this, we identified stronger signatures of selection structuring diversity in older species (**Fig. 3e**). Thus, while there may be species that start on the journey to panmixia, with large population sizes, weak geographic barriers and free recombination (as exemplified by *V. parahaemolyticus*), all such species acquire barriers to recombination due to epistatic selection as they age, and end up with intermediate long-range LD.

An important implication of our findings is that we can gain 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 free genetic exchange 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[37,40]. 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. These suggest similar ecological driving forces and/or adaptive potentials in both species, which might reflect the diverse niches within complex multispecies biofilms in the oral cavity, that could be exploited by evolving cell shape and cell wall properties. 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[39] 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 retrieved all assembled genomes (n=317,542) included in the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org, release 207)[41,42]. We performed genome deduplication

based on BioSample accessions, retaining only the genome with the highest N50 value for each BioSample. Only high-quality genomes (n=268,463) with estimated completeness ≥90% and contamination ≤5%, as assessed by CheckM v1.1.3[45], were retained for subsequent analyses.

The high-quality genomes were assigned to species based on both National Center for Biotechnology Information (NCBI) taxonomy and GTDB classifications. For each species, the earliest sequenced complete genome, or alternatively the genome with the highest N50 value (if no complete genome was available), was designated as the reference genome. All genomes of a species were aligned to the corresponding reference genome using MUMmer v3.23[46] to generate whole-genome alignments, as previously described[36], and only genomes with ≥60% alignment coverage against the reference were retained for subsequent analyses.

After deduplication and quality control, a total of 266,627 high-quality genomes were retained for downstream analyses, and their genome quality metrics and taxonomic classifications are provided in Supplementary Table 1. The whole-genome alignments for each species are publicly available from the Zenodo database (https://zenodo.org/records/17430471).

Representative datasets and clonal groups

While robust estimation of population genetic parameters requires randomly sampled representative datasets, biased repeat sampling (*e.g.*, strains from disease outbreaks) is prevalent in public genomic datasets, especially for pathogenic species. To reduce the effects of repeat sampling, we generated representative genome datasets as previously described[47]. Briefly, for each species, core-genome (regions present in >95% strains) single-nucleotide polymorphisms (SNPs) were identified using SNP-sites v2.5.1[48]. SNPs located in repetitive regions, as identified by TRF v4.07b[49] or BLASTN self-alignment, were excluded from subsequent analyses. Pairwise SNP distances between strains were calculated using snp-dists v0.8.2 (https://github.com/tseemann/snp-dists). Genomes of a species were grouped into complete-linkage clusters (CLCs) based on pairwise SNP distances using the R package ape v5.7.1[50]. For the representative dataset of each species, we used a "relative" threshold, defined as 1% of the median pairwise SNP distance, as the cutoff, and selected a single genome with the highest N50 from each CLC or from singleton genomes as representative genomes.

Estimation of substitution rates and selected recombination parameters rely on clonal groups (CGs) consisting of closely related genomes. We used an "absolute" SNP-distance threshold of 5,000 SNPs (corresponding to the total SNP cutoff per branch for reliable recombination detection[51]) as the cutoff to identify CLCs corresponding to CGs. The assignments of the representative and CG datasets are provided in Supplementary Table 1.

Genomic parameters

We calculated genomic parameters (GPs) across 12 categories for all representative genomes, using the median value across representative genomes to represent each species (Supplementary Table 2 and Table 4). The variability of each parameter, quantified as the standard deviation (SD)-to-median ratio, is shown in Supplementary Fig. 2. Some GP estimates exhibited relatively high variability (SD/median ratio >0.5), which we interpreted as indicative of 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.

Specifically, genome size and GC content were calculated using Quast v5.0.2[52]. We re-annotated representative genomes using Prokka v1.13[53] with default settings, 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 (RM) system genes were retrieved from REBASE[54] and searched against annotated protein sequences to identify RM genes using BLASTp; hits with >50% coverage were considered present as previously described[36]. Repetitive regions of genomes were identified using TRF v4.07b[49] and BLASTN self-alignment as previously described[36]. CRISPR arrays were identified using CRISPRidentify v1.1.0[55]. Phage sequences were identified using PhiSpy v3.7.8[56]. Insertion sequence (IS) elements were identified using ISEScan v1.7.2.3[57]. Default parameters were used for all analyses unless otherwise specified.

Population genetic parameters

We calculated seven categories of population genetic parameters (PGPs) for species with ≥10 representative genomes and for CGs with ≥10 genomes (Supplementary Table 5).

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 alignments. Pairwise non-synonymous divergence (*dN*), synonymous divergence

(*dS*), and *dN/dS* were calculated with PAML v4.10.0[58] 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[59]. For each replicate, estimates were retained only when >1,000 locus pairs were available at the given distance.

h/m

We calculated the ratio of homoplasic and non-homoplasic alleles (*h/m*) using ConSpeciFix v1.3.0[60] 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[61] 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 ($r/m = \varphi_{\text{pool}}/\theta_{\text{pool}}$).

Recombination 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[27]: 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[60].

For each CG of a species, we reconstructed maximum-likelihood (ML) phylogenetic trees using FastTree v 2.1.10[62] based on non-repetitive and non-recombined SNPs as previously described[63]. The isolation year of strains was extracted from the corresponding BioSample information. We used the ML tree and isolation year information to calculate the substitution rate with BactDating

v1.1.0[64] as previously described[47]. Only substitution rate estimates with effective sample sizes (ESS) exceeding 100 were retained, ensuring robust temporal signal and reliable estimation.

Phenotype and ecology data

The phenotype and ecology data were extracted from three sources: Madin *et al*[65], Secaira-Morocho *et al*[66], and the BacDive database (up to August 2025)[67]. 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 a 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). Associations between quantitative and categorical variables were measured with Cramér’s coefficient (V). 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 (Supplementary Table 6).

The phylogenetic signal of all parameters, including GPs, PGPs, and phenotypic and ecological parameters, was quantified using Pagel’s λ and estimated with the R package geiger v2.0.11[50]. Specifically, the fitContinuous function was applied with the model set to "lambda", using phylogenetic trees based on universally distributed marker genes (archaea: 53 genes; bacteria: 120 genes) as implemented in GTDB[41,42], together with the corresponding parameter values as input. Categorical traits were converted into continuous numerical values ordered from low to high, such that identical categorical states were represented by the same numerical value.

Phylogenetically corrected correlation coefficients were estimated for specific parameter pairs shown in Fig. 2 using the R package phylolm v2.6.2[68]. The phylolm function was applied with the

model set to "lambda", incorporating the same GTDB-based phylogenetic trees and corresponding parameter values as input, with the regression formula specified as *parameter A ~ parameter B*.

Principal component analysis (PCA)

PCA was performed using PLINK v1.9[69] 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.*[39], 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. SNPs in the top 1% of loading values for the first principal component were defined as PC1-related SNPs, representing the genetic basis of major population structure, whereas the remaining SNPs were treated as genomic background.

SNP annotation and gene enrichment analysis

Synonymous and non-synonymous SNPs were annotated using SnpEff v4.3t[70], based on the re-annotated reference genome. Genes associated with PC1-related SNPs were extracted for subsequent analysis. Functional enrichment was performed using the R package ClusterProfiler v4.2.2[71].

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 PC1-related SNPs were used for phylogenetic tree construction using FastTree v2.1.10[62].

Data availability

All genome sequences are available from the Genome Taxonomy Database (GTDB; https://gtdb.ecogenomic.org, release 207). Essential data required for replication, including whole-genome alignments, SNP matrices, curated metadata, and calculated results, are available at www.genomap.cn and Zenodo (https://zenodo.org/records/17430471), and are also provided in Supplementary Tables. Linux command-line scripts used to calculate genomic and population genetic parameters are provided in the Supplementary Text.

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).
3. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333-351 (2016).
4. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet*. **22**, 572-587 (2021).
5. Blanchet, S., Prunier, J. G. & De Kort, H. Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends Genet.* **33**, 579-580 (2017).
6. Leigh, D. M. *et al.* Opportunities and challenges of macrogenetic studies. *Nat. Rev. Genet*.

22, 791-807 (2021).

7. Manel, S. *et al.* Global determinants of freshwater and marine fish genetic diversity. *Nat. Commun*. **11**, 692 (2020).
8. Theodoridis, S. *et al.* Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nat. Commun.* **11**, 2557 (2020).
9. Vellend, M. & Geber, M. A. Connections between species diversity and genetic diversity. *Ecology Letters* **8**, 767-781 (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* **353**, 1532-1535 (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**, 195-205 (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* **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**, 362-378 (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**, 2740-2745 (2001).
21. Weigand, M. R. *et al.* The history of Bordetella pertussis genome evolution includes structural rearrangement. *J. Bacteriol.* **199**, 10.1128/jb. 00806-16 (2017).
22. Zafar, H. & Saier, M. H. Gut Bacteroides species in health and disease. *Gut Microbes* **13**, 1848158 (2021).
23. Celli, J. The Intracellular Life Cycle of Brucella spp . *Microbiol. Spectr.* **7**, 10.1128/microbiolspec. bai-0006-2019 (2019).
24. Teng, W., Liao, B., Chen, M. & Shu, W. Genomic Legacies of Ancient Adaptation Illuminate GC-Content Evolution in Bacteria. *Microbiol. Spectr.* **11**, e02145-22 (2023).
25. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Molecular Biology Reports* **49**, 539-565. (2022).
26. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nat. Rev. Genet.* **17**, 422–433 (2016).
27. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **11**, 1–18 (2015).

28. Fuhrman, J. A. *et al.* A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7774-7778 (2008).
29. Andam, C. P. *et al.* A latitudinal diversity gradient in terrestrial bacteria of the genus Streptomyces. *mBio* **7**, 10.1128/mbio. 02200-15 (2016).
30. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115 (2010).
31. 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).
32. Yang, C., Cui, Y., Didelot, X., Yang, R. & Falush, D. Why panmictic bacteria are rare. *bioRxiv* (2018): 385336.
33. 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**, 256 (2012).
34. Thorpe, H. A. *et al.* Repeated out-of-Africa expansions of Helicobacter pylori driven by replacement of deleterious mutations. *Nat. Commun.* **13**, 6842 (2022).
35. Kinjo, Y. *et al.* Enhanced Mutation Rate, Relaxed Selection, and the 'domino Effect' are associated with Gene Loss in Blattabacterium, A Cockroach Endosymbiont. *Mol. Biol. Evol.* **38**, 3820-3831 (2021).
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)
39. Liu, S., Svensson, S. L. & Falush, D. A bacterial ecocline in Klebsiella pneumoniae may explain its backboned phylogeny. *PLoS Biol.* **24**, e3003672 (2026).
40. Tourrette, E. *et al.* An ancient ecospecies of Helicobacter pylori. *Nature* **635**, 178–185 (2024).
41. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).
42. 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).
43. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25**, 39-67 (2001).
44. 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).
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**, 10.3.1-10.3.18. (2003).
47. Yang, C. *et al.* Wave succession in the pandemic clone of Vibrio parahaemolyticus driven by gene loss. *Nat. Ecol. Evol.* (2025).

48. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
49. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
50. Pennell, M. W. *et al.* Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216-2218 (2014).
51. 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).
52. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
53. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
54. 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**, D629-D630 (2023).
55. Mitrofanov, A. *et al.* CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res.* **49**, e20-e20 (2021).
56. 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**, e126-e126 (2012).
57. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, 3340-3347 (2017).
58. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).
59. Barrett, J. C. Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harb. Protoc.* (2009): pdb-ip71.
60. Bobay, L. M., Ellis, B. S. H. & Ochman, H. ConSpeciFix: Classifying prokaryotic species based on gene flow. *Bioinformatics* **34**, 3738-3740 (2018).
61. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
62. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).
63. 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).
64. 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**, e134-e134 (2018).
65. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Sci. Data* **7**, 170 (2020).
66. 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).
67. Reimer, L. C. *et al.* BacDive in 2022: The knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.* **50**, D741-D746 (2022).

68. Tung Ho, L. S. & Ané, C. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.* **63**, 397-408 (2014).
69. 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).
70. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80-92 (2012).
71. 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.32350710791 and No. 32170640 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.

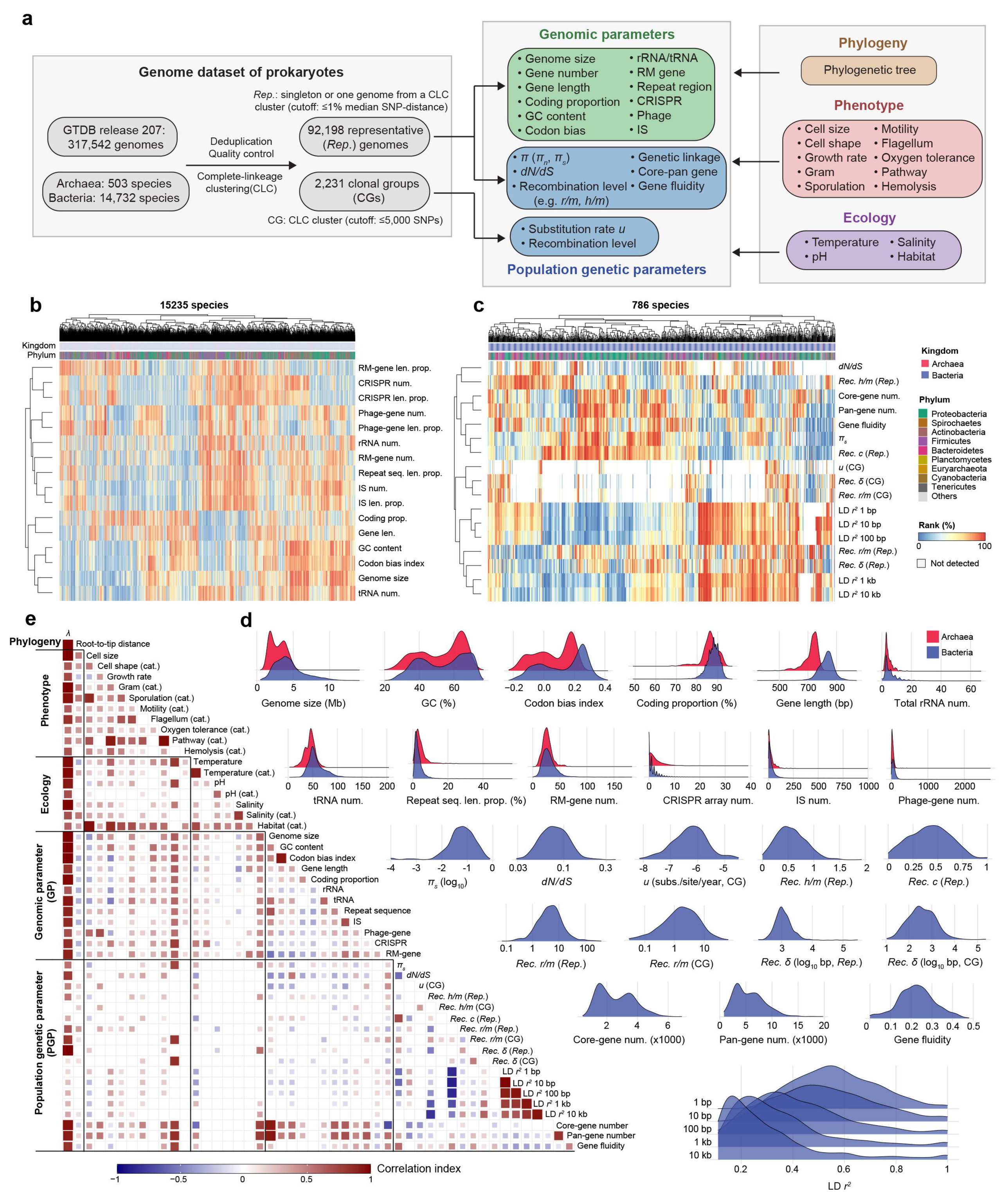
Figure legends

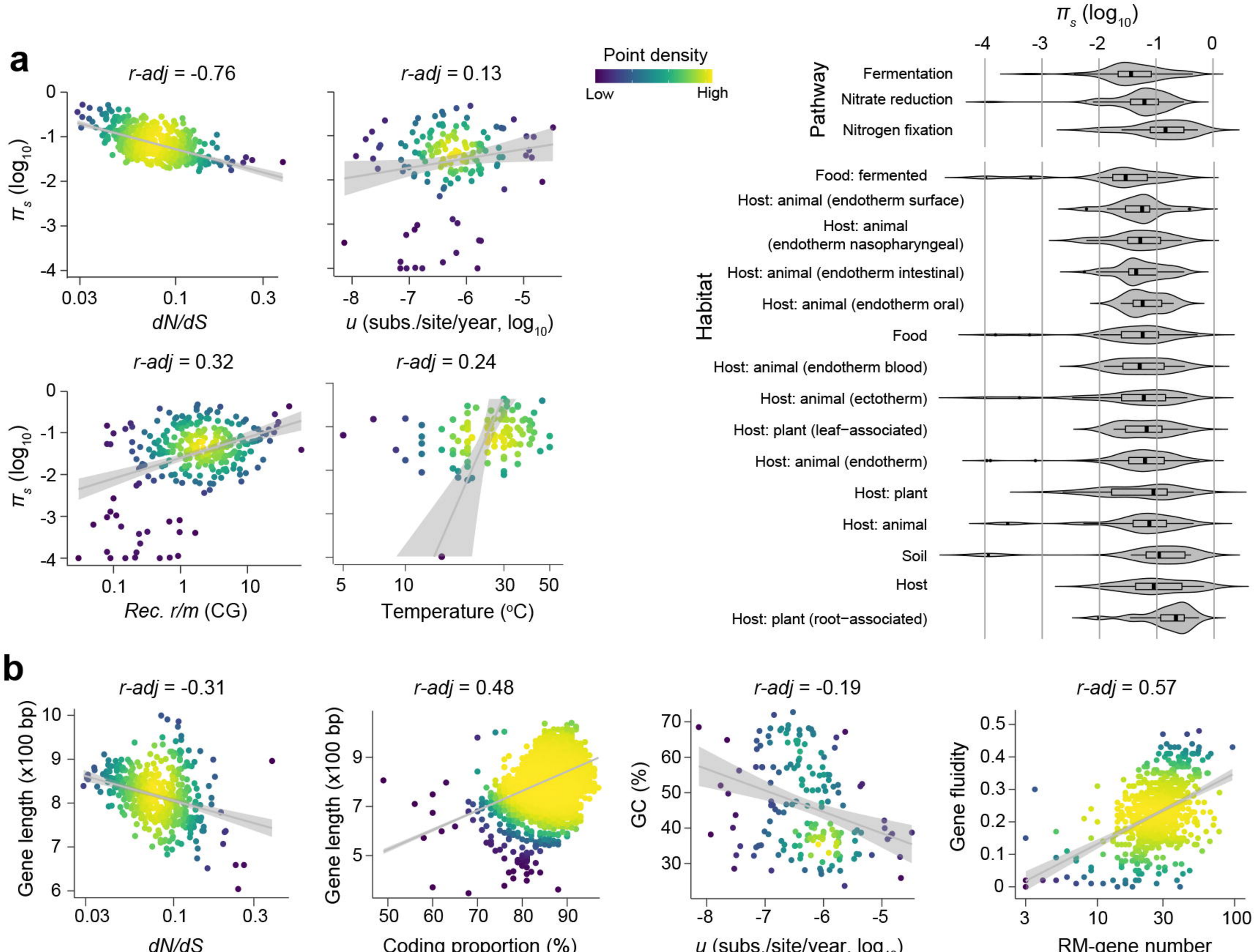
Figure 1. Overview of the MAP. **a)** Study flowchart. **b,c)** Heatmaps showing the rank percentages of genomic parameters (GPs) and population genetic parameters (PGPs) for each species. Bars at the top indicate the kingdom and phylum, while heatmap colors represent rank percentages. **d)** Distribution of different GPs and PGPs. 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-related 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.

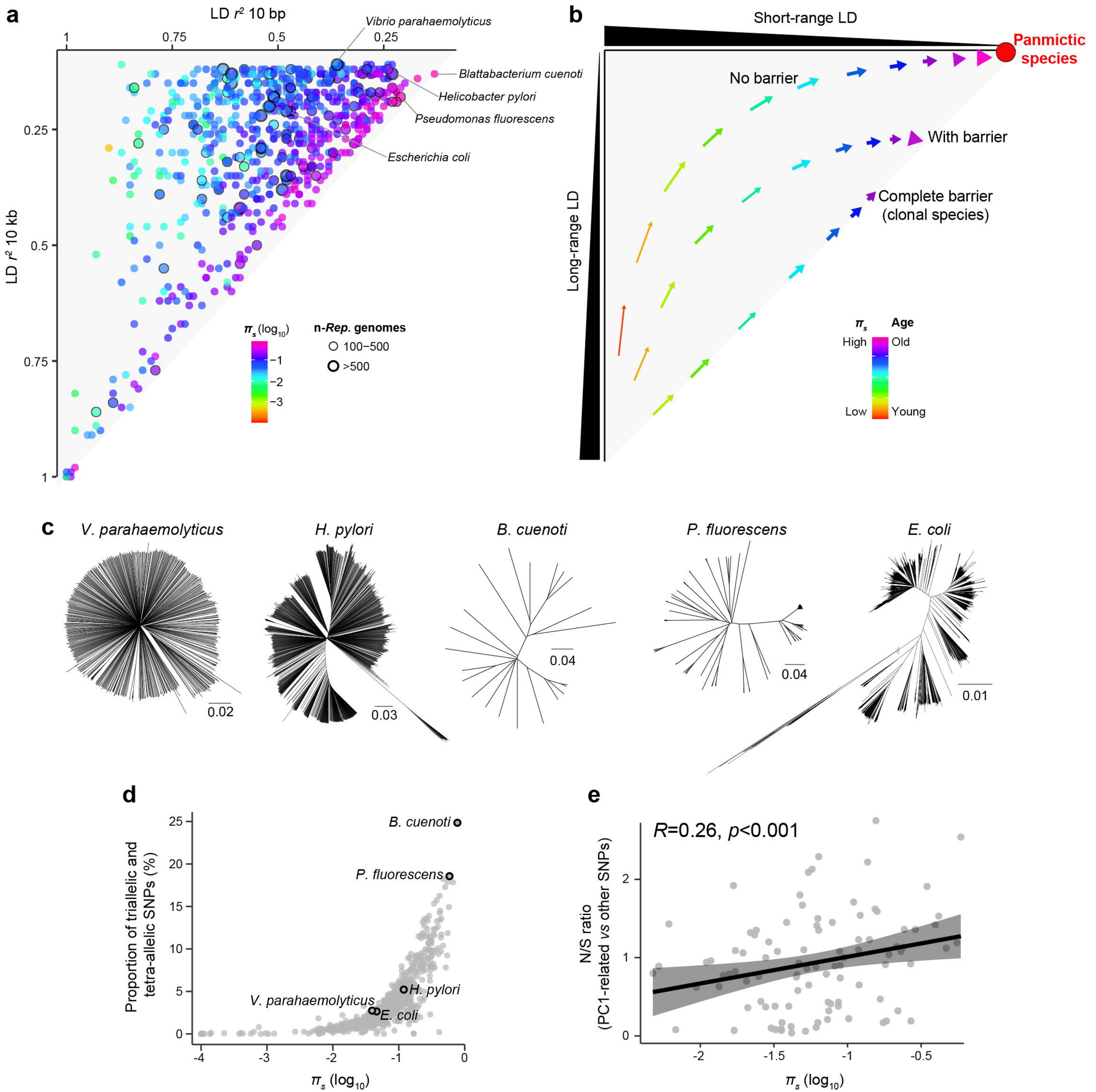
Figure 2. Correlations related to genetic diversity and underexplored associations. a) Correlations associated with genetic diversity. b) Underexplored correlations revealed by MAP. *r-adj* represents the correlation coefficient after adjustment for phylogenetic signal. Point colors represent point density. 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.

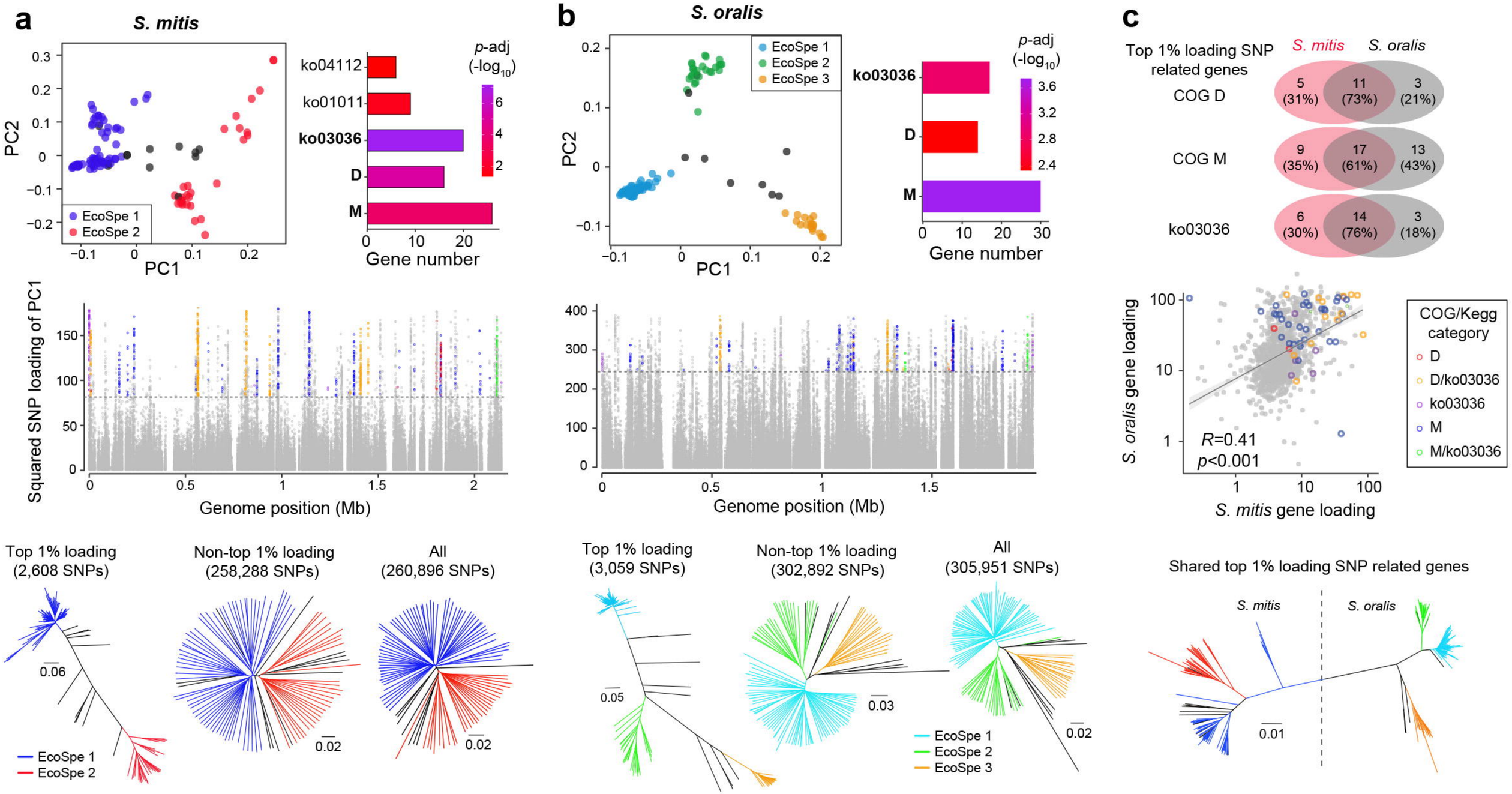
Figure 3. Overview of prokaryotic genetic diversity. a) Relationships between short-range (10 bp), long-range LD (10 kb) and genetic diversity. Point colors represent genetic diversity levels. Species with ≥100 representative genomes are outlined in black. **b)** Conceptual model illustrating the evolution 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. Panmictic species denote freely recombining and highly diverse species with minimal LD at all distances. **c)** Phylogenetic trees of five representative species. **d)** Correlation between genetic diversity and the proportion of triallelic and tetra-allelic SNPs. Five representative species are highlighted with black outlines. **e)** Correlation between genetic diversity and the N/S ratio, defined as the ratio of non-synonymous to synonymous sites for SNPs in the top 1% of loadings on PC1 (PC1-related SNPs) relative to the remaining SNPs.

Figure 4. Convergent ecospecies. a,b) Top: Principal component analysis (left; point colors indicate inferred ecospecies) and functional enrichment of genes associated with PC1-related SNPs (top 1% by loading values; right) in *Streptococcus mitis* and *S. oralis*. Middle: Genome-wide distribution of SNP loadings on PC1. Point colors represent COG or KEGG categories. Dashed lines indicate the top 1% loading cutoff. Bottom: Phylogenetic trees based on PC1-related SNPs, the remaining SNPs and all SNPs. Branch colors indicate inferred ecospecies. **c)** Top: Shared homologous genes associated with PC1-related SNPs 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 PC1-related SNPs.


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
λ
Phylogeny
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
Correlation index
−1
−0.5
0
0.5
1
d
Archaea
Bacteria
Genome size (Mb)
GC (%)
Codon bias index
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 (Rep.)
Rec. c (Rep.)
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 = -0.76
r-adj = 0.13
r-adj = 0.32
r-adj = 0.24
Point density
Low
High
π_s ($\log_{10}$)
dN/dS
u (subs./site/year, $\log_{10}$)
Rec. r/m (CG)
Temperature (°C)
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)
b
r-adj = -0.31
r-adj = 0.48
r-adj = -0.19
r-adj = 0.57
Gene length (x100 bp)
GC (%)
Gene fluidity
Coding proportion (%)
RM-gene number


a
LD r² 10 bp
1
0.75
0.5
0.25
Vibrio parahaemolyticus
Blattabacterium cuenoti
Helicobacter pylori
Pseudomonas fluorescens
Escherichia coli
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
Complete barrier (clonal species)
Long-range LD
πs
Age
High
Old
Low
Young
c
V. parahaemolyticus
H. pylori
B. cuenoti
P. fluorescens
E. coli
0.02
0.03
0.04
0.04
0.01
d
Proportion of triallelic and tetra-allelic SNPs (%)
B. cuenoti
P. fluorescens
H. pylori
V. parahaemolyticus
E. coli
πs (log10)
e
R=0.26, p<0.001
N/S ratio (PC1-related vs other SNPs)
πs (log10)


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)
Top 1% loading (2,608 SNPs)
Non-top 1% loading (258,288 SNPs)
All (260,896 SNPs)
0.06
0.02
0.02
b
S. oralis
EcoSpe 1
EcoSpe 2
EcoSpe 3
ko03036
D
M
Top 1% loading (3,059 SNPs)
Non-top 1% loading (302,892 SNPs)
All (305,951 SNPs)
0.05
0.03
0.02
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
Shared top 1% loading SNP related genes
0.01