Macrogenetic atlas of prokaryotes

Chao Yang[1,*], Hao Huang[1], Naike Wang[1], Xavier Didelot[2], Ruifu Yang[3], Yujun Cui[3], Daniel Falush[1,*]

[1] The Center for Microbes, Development and Health, CAS Key Laboratory of Molecular Virology and Immunology, Shanghai Institute of Immunity and Infection, Chinese Academy of Sciences, Shanghai, China;

[2] School of Life Sciences and Department of Statistics, University of Warwick, UK

[3] State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, China;

* Corresponding authors:

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

Abstract

Macrogenetics investigates the patterns and predictors of intraspecific genetic variation across diverse taxa, offering a framework to address longstanding evolutionary hypotheses. Here, we present a comprehensive macrogenetic atlas of prokaryotes (MAP), integrating genomic data (summarized by 30 parameters in 12 categories) from 15,235 prokaryotic species and population genetic data (35 parameters in 7 categories) from 786 species with phylogenetic, phenotypic, and ecological data. We illustrate the power and potential of the MAP by addressing a series of fundamental questions. For example, we provide evidence that genome size in prokaryotes depends on a balance between deletion bias and the beneficial effects of additional genes. Moreover, we show that long-range and short-range linkage disequilibrium represent quite different features of evolutionary dynamics and contribute independently to genetic diversity. Notably, we show that as within-species diversity increases, selection becomes an increasingly strong force structuring diversity, for example creating convergent ecospecies structures in *Streptococcus mitis* and *S. oralis*. Overall, our atlas represents a widely applicable resource and offers novel insights into the drivers of macroevolution and the life-cycle of prokaryotic species.

Introduction

Over the past 50 years, population genetics has greatly advanced our understanding of fundamental questions regarding the nature of genetic variation and evolution, while providing essential underpinnings for fields such as evolutionary biology and ecology[1]. More recently, the rapid advent of high-throughput sequencing technologies has dramatically propelled this field forward by enabling the generation of large-scale genomic data with unprecedented scope[2,3]. Classical population genetics focuses primarily on genetic diversity and evolutionary forces within individual species. Now, with the ongoing accumulation of data across many taxa, the time is ripe for a novel view—macrogenetics—which embraces broad taxonomic, spatial or temporal scales across dozens to thousands of species[4,5].

Macrogenetics holds significant promise for addressing longstanding evolutionary hypotheses. For instance, recent macrogenetic studies have uncovered a positive correlation between species diversity within a community and intrapopulation genetic diversity in fish and mammals[6,7], providing evidence for the species-genetic diversity correlation (SGDC) concept[8]. Additionally, macrogenetic studies also offer novel insights into what determines genetic diversity[9], revealing the species traits (fecundity and propagule size) underlying genetic diversity[10] and latitudinal gradients in genetic diversity[11].

Prokaryotes make up more biomass than all organisms other than plants combined[12]. Furthermore, a single prokaryotic species can have lineages within it that specialize to multiple distinct niches, for example differing in host range, pathogenicity and nutrient sources, the classic example being *Escherichia coli*[13]. This feature contrasts with eukaryotes where there is typically limited genetic differentiation and specialization within species. This difference raises a number of specific questions such as how are ecological niches partitioned between prokaryotic species, and how does phenotypic and niche diversity affect the genomic and population genetic characteristics of each species. These differences also mean that the process of speciation is likely to be driven by a different set of driving forces, which are collectively less well understood than those in eukaryotes[14].

Here we present a comprehensive macrogenetic atlas of prokaryotes, integrating large-scale genomic and population genetic data with phylogenetic, phenotypic, and ecological metadata. This atlas is a widely applicable resource, and we demonstrate its utility in addressing long-standing evolutionary hypotheses and offering novel insights into the process of speciation.

Results

Overview of macrogenetic atlas of prokaryotes

We analyzed 317,542 assembled genomes across 15,235 species (503 archaeal and 14,732 bacterial species; NCBI taxonomy). Following deduplication and quality control, a representative (*Rep.*) dataset for each species was constructed to minimize biases from repeat sampling (see Methods). To incorporate microevolution data, we also constructed one or more clonal group (CG) datasets for each species, consisting of sets of closely related genomes.

For genomic parameters (GP), we used the median value of *Rep.* genomes as the species estimate. To estimate population genetic parameters (PGP), we performed random sampling of ten *Rep.* genomes for each species across 10 independent replications, accounting for variable sample sizes. The median PGP value from these replications was used as the species estimate. The CG dataset was exclusively used to estimate mutation rates and recombination levels. In total, we obtained 30 parameters from 12 categories of GP for all 15,235 species and 35 parameters from 7 categories of PGP for 786 species (**Fig. 1a,** Supplementary Tables 1 and 2). Moreover, we correlated these parameters with phylogenetic trees, as well as phenotypic (ten categories) and ecological (four categories) data retrieved from public databases[15–17] .

We refer to the integrated GP and PGP data as the macrogenetic atlas of prokaryotes (MAP). The parameter values were transformed into rank percentages for each species to make them distribution-agnostic and to provide a comprehensive overview (**Fig. 1b,c**). We quantified the correlations among various categories (using r for quantitative data or η for categorical data), as well as their relationships with the phylogenetic tree (Pagel's λ), demonstrating prevalent correlations and strong phylogenetic signals for most parameters (**Fig. 1d**). The distribution of different parameters was assessed **(Fig. 1e)**, showing similar GP distributions for archaea and bacteria, with the exception that archaeal species tend to have shorter gene lengths. Most parameters exhibited normal-like, left-skewed, or right-skewed distributions, while bimodal distributions were rarely identified, with the notable exceptions of GC content and codon bias index (CBI).

Interestingly, we observed that some "well-studied" species (with >100 *Rep*. genomes) exhibit "extreme" parameter values. For instance, the human pathogen *Helicobacter pylori* displays an exceptionally high proportion of restriction-modification (RM) system genes, comprising 5.1% of its genome, ranking first among 15,235 species. Additionally, we identified "extreme" genera. For example, the cell-wall-deficient helical *Spiroplasma* bacterial genus has an exceptionally low average gene length of 278 bp, with 7 of the lowest 10 species in gene length belonging to this genus.

Determinants of genomic characteristics

The MAP enabled a comprehensive investigation of fundamental questions, such as what determines genome size. To explore this, we extracted several genomic parameter (genome size, gene length and GC content) related network (r or $\eta \geq 0.3$, **Fig. 2a**) and adjusted for phylogenetic signals in the correlations (quantified by *r-adj*), due to the significant phylogenetic signals observed across most parameters. Following adjustment, the correlation coefficients for most parameter pairs decreased, with a notable reduction in the correlation between genome size and GC content ($r = 0.5$, *r-adj* = 0.1).

Genome size

Genome replication is costly, so natural selection will remove unnecessary sequence. In support of this prediction, we have found a negative correlation between genome size and growth rate (*r-adj* = -0.13, **Fig. 2b**). Genome streamlining theories, which have found good empirical support in eukaryotes, therefore predict that species where selection is more effective should have shorter,

more compact genomes[18,19]. In prokaryotes, a convenient way to measure the effectiveness of selection is using *dN/dS* (the ratio of non-synonymous to synonymous substitution rates), with lower *dN/dS* or high *dS/dN* values found in species where purifying selection is more effective.

Previous work in bacteria has found that *dN/dS* is negatively correlated with genome size, which goes against the expectation under the "streamlining" theory, since weaker selection should cause both a higher *dN/dS* and a higher genome size[19]. It has therefore been proposed that in prokaryotes stronger selection allows a larger number of beneficial genes to remain in the genome, in the face of ubiquitous deletion bias[19]. Other authors have found a negative correlation between mutation rate (*u*) and genome size[20]. We validated the previous observations of negative correlations between *dN/dS* and genome size (*r-adj* = -0.22), as well as between *u* and genome size (*r-adj* = -0.26). Furthermore, we showed that the compound parameter *dS/dN*×*u* (*r-adj* = -0.59) was most predictive of genome size (**Fig. 2b**). This observation provides additional support for a "balance" model, with the two factors acting in opposition to each other to determine the overall genome size.

We have also found several other factors influencing genome size including the habitat, temperature and oxygen tolerance. For example, aerobic species living in soil and in low temperatures tend to have larger genomes. Some of these factors, such as gene number, core-/pan-genes, RM-genes, tRNA counts, and cell size are likely outcome factors of genome size. Other factors correlated with genome size are metabolic pathways, cell shape and sporulation (**Fig. 2b**).

Gene length

Similar reasoning to the determinants of genome size should also predict that species with more effective selection should also have longer genes. Gene length is more strongly correlated with *dN/dS* (*r-adj* = -0.31) than *u* (*r-adj* = -0.14) or the compound parameter *dS/dN*×*u* (*r-adj* = -0.29, **Fig 2c**), implying that selection rather than deletion bias is a driver of gene length. Moreover, there is a strong positive correlation between gene length and coding proportion (*r-adj* = 0.48, **Fig. 2c**), which could be explained by non-coding regulatory regions being of similar length for long and short genes.

GC content

GC content is highly variable in bacteria despite empirical evidence of a universal mutation bias towards AT[21]. We identified a significant negative correlation between mutation rate and GC content (*r-adj* = -0.19, **Fig. 2d**), suggesting that a high input of A and Ts through mutation acts to lower the GC content, while the factors counteracting this mutation bias need time to act.

Habitat emerged as another potential causal factor, with soil-isolated species exhibiting higher GC content. In contrast, GC content had a significant positive effect on codon bias (*r-adj* = 0.84). Furthermore, we observed that aerobic species with rod-shaped cells and monotrichous or polar flagella tend to have an elevated GC content (**Fig. 2d**).

Determinants of population genetic characteristics

Genetic diversity

Classical population genetic theory predicts that in a well-mixed population, synonymous nucleotide diversity π_s should be proportional to the product on the mutation rate u and the effective population size N_e[9]. In practice, we do not have a direct measure of N_e for any species, however in eukaryotes, the negative correlation between propagule size or fecundity and π_s[10] is likely to be driven by higher N_e of fecund organisms and those with small propagules. In our atlas, there is a negative correlation (*r-adj* = -0.14) between cell size and within-species π_s (**Fig. 3a,b**), consistent with the pattern seen in eukaryotes. *dN/dS* can be used as a proxy for N_e since selection decreases both[14], and we find a strong negative correlation between *dN/dS* and π_s (*r-adj* = -0.76), consistent with theory. There is also a weak correlation between the mutation rate u and π_s (*r-adj* = 0.13). However, the relationships between other parameters is more complex. Specifically, there is an unexpected negative correlation between the compound parameter $dS/dN \times u$ and π_s (*r-adj* = -0.73).

In terms of potential ecological factors **(Fig. 3b)**, the habitat with the highest within-species π_s are plant roots, which might reflect the stability of these habitats and the large bacterial population sizes associated with them. By contrast, fermented food has the lowest diversity, which might reflect artificial selection but also the bottlenecks associated with the fermentation cycle. Concordant with this, species involved in fermentation have lower diversity than those that fix nitrogen. There is also a positive correlation with temperature (*r-adj* = 0.24), similar to the higher genetic diversity observed in eukaryotic species living in tropical regions[11], which might reflect warmer habitats supporting larger populations of both prokaryotes and eukaryotes, as well as seasonal bottlecks in colder climbs.

According to population genetic theory, linkage disequilibrium (LD) values within a species are determined by N_e and the recombination rate[22]. In species which maintain large well-mixed populations for long periods, recombination will break down LD, meaning that there should be low LD at all genetic distances and high π_s. We have examined correlations between π_s and LD at different genetic distances. Concordant with theory, there is a strong negative correlation between short-range LD (10 bp) and π_s (*r-adj* = -0.58). However, while long-range LD (10 kb) and short-range LD are highly correlated (*r-adj* = 0.68), the correlation between long-range LD and π_s is not significant. Furthermore, for species with similar levels of short-range LD, there is a pronounced positive correlation between long-range LD and π_s (**Fig. 4a**). Thus long-range and short-range LD reflect evolutionary dynamics that contribute independently in opposite directions to genetic diversity.

Recombination is expected to positively influence genetic diversity, for example by reducing the fraction of the genome influenced by selective sweeps. We identified significant positive correlations between π_s and several recombination-related parameters (**Fig. 3b**): the proportion of genome sites affected by recombination (*Rec. c*[23], *r-adj* = 0.72), the ratio of effects of recombination and mutation (*r/m*[24], *r-adj* = 0.32) and the ratio of homoplasic and non-homoplasic alleles (*h/m*[25], *r-adj* = 0.16). The correlations between *c*, *h/m* and π_s are more likely mediated by LD, given the stronger correlation of these parameters with LD than with π_s (*c* vs r^2 10 bp: *r-adj* = 0.89, *h/m* vs r^2 10 kb: *r-adj* = 0.73). However, the correlation between *r/m* and π_s is stronger than

with LD (r^2 10 bp: *r-adj* = 0.2, r^2 10 kb: *r-adj* = 0.19), suggesting that LD statistics do not capture the full effect of recombination on diversity.

Mutation rate

The determinants of mutation rate remain poorly understood. One potentially revealing explanation, the drift-barrier hypothesis, suggests that selection minimizes mutation rates to enhance replication fidelity, with selection efficiency constrained by genetic drift[26]. However, we did not observe a significant correlation between *dN/dS* and *u* ($p > 0.05$). Instead, temperature emerged as a potential causal factor (*r-adj* = 0.23), with species in warmer environments showing higher mutation rates (**Fig. 3c**).

Gene fluidity

Lifestyle, effective population size, and genome size have been reported as major determinants of prokaryotic pangenomes[27–29]. Consistent with previous studies, we identified significant correlations between habitat (η = 0.53), *dN/dS* (*r-adj* = 0.53), genome size (*r-adj* = 0.24) and gene fluidity (**Fig. 3d**), measured by the ratio of unique gene families to the total gene families in genome pairs[28]. Gene fluidity positively affects pan-gene number (*r-adj* = 0.63), but has a weak effect on core-gene number (*r* = -0.1).

We found a strong positive correlation between the number of RM genes (*r-adj* = 0.57) and gene fluidity. Given the role of RM genes in regulating gene transfer[30], we propose RM genes as a putative causal factor for gene fluidity. Moreover, the number of mobile genetic elements, such as phages (*r-adj* = 0.29) and insertion sequence elements (ISE, *r-adj* = 0.2), are also putative causal factors due to their role in promoting gene transfer.

Other factors, including cell shape, growth rate, metabolic pathways and oxygen tolerance, were also significantly correlated with gene fluidity, though the causal direction remain unclear. Notably, a strong connection was found between cell shape and gene fluidity, with rod-shaped species exhibiting more than a two-fold higher gene fluidity than spherical-shaped species. Furthermore, the correlations between *u*, *r/m*, and gene fluidity suggest shared drivers and molecular mechanisms that influence both core and accessory genomic variations, as well as homologous and non-homologous recombination.

Absence of panmictic species

Many bacterial species disperse effectively around the globe and have enormous population sizes. Accordingly, we have observed several species with very high genetic diversity, comparable to that in most diverse eukaryotes[31], such as *Pseudomonas fluorescens* (**Fig. 4a,b**). Classical population genetic theory suggests that these species should be panmictic, characterized by low LD at all genetic distances. However, while the most diverse species do have low short-range LD, consistent with high recombination rates, the diverse species are all highly structured (**Fig. 4a,b**), reflected by their moderate long-range LD. Consequently, there is a paradoxical absence of panmictic species.

An example of a well-mixed species with low long-range LD is *Vibrio parahaemolyticus*, which has moderate diversity. Previous analysis has suggested that diversity of *V. parhaemolyticus* is moderate because its population size has expanded only recently[32] and that if population size were maintained, equilibrium diversity would be much higher, and short-range LD would become lower, which makes it a prime candidate to evolve into a panmictic species. However, *V. parahaemolyticus* is also in the initial stages of functional diversification into ecospecies, which are highly differentiated from each other, but only in a small fraction of the genome[33,34]. If long-range LD increases over time due to an expansion of functional diversification across the genome, then *V. parahaemolyticus* would not evolve into a panmictic species but instead come to resemble the structured high diversity species that we have observed.

Our hypothesis is therefore that panmictic bacteria are absent because in species that have maintained large population sizes, selection will progressively create long-range LD, due to inevitable functional diversification. To test this hypothesis, we examined the genetic basis of population structure within bacteria. Specifically, we used principal component analysis (PCA) to identify the major population structure in each species. We hypothesized that this structure should show stronger signatures of selection in more diverse species. As a simple measure of selection, we looked for an enrichment of non-synonymous sites amongst the SNPs most strongly associated with particular PC components, as measured by SNP loadings. We expect selection to be evident when species have maintained their current population size for long time periods, which should result in similar values of short- and long-range LD. Accordingly, we introduce an LD saturation statistic, defined by the ratio of long-range to short-range LD. As predicted by the hypothesis, we observed a strong enrichment of such non-synonymous sites amongst species with high saturation levels (**Fig. 4c**), which we tested statistically by computing the correlation between enrichment and saturation when the latter is >0.5 (zones 3 and 4 in **Fig. 4a,c**). The correlation is strongest for species with lower levels of short-range LD (zone 4), indicating selection is more effective for highly recombining species (**Fig. 4c**), supporting our hypothesis. The absence of correlation for species with low saturation (zones 1 and 2 in **Fig. 4a,c**) shows that selection only becomes evident in patterns of genome wide diversity for species that have maintained their current population size for long periods.

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[33,34], implying high differentiation in specific regions and high gene flow elsewhere (**Fig. 4d,e**). Furthermore, there is strong enrichment of non-synonymous SNPs and genes of particular categories within SNPs 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 congruence 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. 4f**).

To characterize these putative ecospecies, we constructed phylogenetic trees based on SNPs with extreme loadings (**Fig. 4d,e**). We found two putative ecospecies for *S. mitis* and three for *S. oralis*, which also correspond to distinct clusters in strain PCA loadings. 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. 4f**). Finally, we constructed phylogenetic trees ased on remaining 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*[33,34]. 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

The MAP resource should be useful to any scientist interested in a particular microbe, by providing a simple summary of genomic, population genetic and other parameters and allowing easy comparison between all available prokaryotic species. It also represents a resource for macrogenomic and population genetic analysis, including raw data and non-redundant datasets (see Data availability). We have illustrated the power and potential of MAP by addressing a series of fundamental questions, and by comparing our results to those obtained from eukaryotes, which remain better studied. For example, we provide evidence that genome size depends on a balance between deletion bias and the beneficial effects of additional genes[19]. Secondly, we show that long-range and short-range LD represent quite different features of evolutionary dynamics and contribute independently to genetic diversity. Thirdly, we have uncovered a surprisingly strong relationship between cell shape and genome fluidity, with rod-shaped species having more than two-fold higher fluidity than spherical ones.

Our analyses and resources are available based on two taxonomic classifications, namely NCBI and Genome Taxonomy Database (GTDB)[35,36]. The problem of bacterial species definitions remains unsolved and large databases inevitably contain errors[37,38]. 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. 1). In many cases, these species do not correspond to coherent evolutionary units. In particular, the most diverse species, such as *P. flourescens*, 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 (Supplementary Fig. 1). Therefore GTDB is not especially well-suited for macrogenomic analysis, especially for highly diverse species, and we presented resulted based on NCBI classification. We nevertheless provide an alternative version of parameter calculation and correlation analysis results based on GTDB (Supplementary Tables 3-4), and show a strong correlation between parameter relationships derived from two taxonomies (r = 0.97, Supplementary Fig. 2), indicating that the great majority of our results are robust to species definitions.

Our population genetic analysis of hundreds of species gives insight into the life-cycle of prokaryotic species and the speciation process. We noticed a paradoxical feature, namely that despite there being several 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[32]. Our resolution of the paradox is that as within-species diversity increases, selection becomes an increasingly strong force structuring diversity.

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*, which are well-mixed species with moderate diversity[33,34]. In these species, ecospecies structure is restricted to a small fraction of the genome. Here we have identified two additional species with ecospecies-like structure, namely *S. mitis* and *S. oralis*. What is particularly fascinating is the convergent nature of the ecospecies structure, with the same cell division and biogenesis associated genome categories and genes involved. These species are more diverse and correspondingly, their ecospecies show evidence for some level of differentiation genome-wide, which 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. flourescens* resists any immediate classification. Characterizing the full range of bacterial 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)[35,36]. We performed deduplications based on the Biosample accessions, retaining only the genome with highest N50 for each Biosample. Genomes with completeness <90% or contamination >5% (estimated by CheckM[40]) were excluded.

The remaining genomes (n=268,463) were grouped into species based on National Center for Biotechnology Information (NCBI) taxonomy (taxid) 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[41] 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 sampling (*e.g.*, strains from disease outbreaks) is common for public datasets, especially for pathogenic species. To minimize 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[42], as previously described[43]. SNPs located in repetitive regions identified by TRF v4.07b[44] 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.

Mutation rate and 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[45]) as the cutoff to identify CLCs representing CGs.

Genomic parameters

We calculated 12 categories of genomic parameters for all the representative genomes, using the median value of representative genomes to represent each species. Genome size and GC content were calculated using Quast v5.0.2[46]. We re-annotated representative genomes using Prokka v1.13[47] and extracted or calculated the gene number and coding proportion 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 annotation results.

The standard protein sequences of restriction-modification (R-M) system genes were retrieved from REBASE[48] and used for searching against annotated protein sequences to identify R-M genes using BLASTP, as previously described[49]. Repetitive regions of genomes were identified

using TRF v4.07b[44] and BLASTN self-search. CRISPR arrays were identified using CRISPRidentify v1.1.0[50]. Phage sequences were identified using PhiSpy v3.7.8[51]. Insertion sequence (IS) elements were identified using ISEScan v1.7.2.3[52]. Default parameters were used for these analysis.

Population genetics parameters

We calculated 7 categories of population genetics parameters for species with ≥10 representative genomes and for CGs with ≥10 genomes. To account for the varying number of representative genomes per species, which can introduce bias in parameter estimation, we performed random sampling of 10 genomes from all representative genomes. The random sampling was independently replicated 10 times, and the median value of the parameters from the replications with ≥5 successful outcomes was used to represent each species.

π_s and *dN/dS*

For each species, the coding sequence alignment 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[53] using the yn00 algorithm. The mean pairwise value of *dS* was used to represent the synonymous nucleotide diversity (π_s).

Recombination

Recombination parameters were calculated for each species using three different methods. For each species, we calculated five parameters using mcorr[23] based on all the representative genomes: 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 ($r/m = \varphi_{pool}/\theta_{pool}$).

We calculated the ratio of homoplasic and non-homoplasic alleles (h/m) using ConSpeciFix v1.3.0[25] based on randomly sampled representative genomes, as well as based on all the genomes for each CG. The median h/m value obtained from random sampling was used to represent each species.

Using all the genomes for each CG, we calculated the parameters of relative rate of recombination to mutation (R/θ), mean divergence between donor and recipient (v), δ and r/m using ClonalFrameML v1.12[24].

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[54] based on randomly sampled representative genomes. The r^2 values were retained only when the number of loci pairs within the given distance exceeded 1,000. We also calculated the LD satuation level, defined as the ratio of r^2 between different distances (e.g. r^2 at 10 kb / r^2 at 10 bp).

Core-genome, pan-genome and gene fluidity

For each species, the number of core- and pan-genes was calculated using Panaroo v1.2.8[55] based on re-annotated results of randomly sampled representative genomes. Gene fluidity was measured by the mean pairwise ratio between the number of genes not shared and the total number of genes.

Mutation rate

For each CG of a species, we reconstructed maximum-likelihood (ML) phylogenetic trees using FastTree v 2.1.10[56] based on non-repetitive and non-recombinational SNPs, as previously described[43]. 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[57]. Only mutation rates with effective sample sizes (ESS) >100 were retained. The median mutation rate of CGs was used to represent each species.

Phenotype and ecology data

The phenotype and ecology data were extracted from three sources: BacDive database[15], Madin *et al*[16] and Secaira-Morocho *et al*[17]. The median value of quantitative data and the major status of category data of multiple strains was used to represent each species.

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 ($p < 0.05$ 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[58]. The phylogenetic trees based on universally distributed genes (archaea: 53 genes, bacteria: 120 genes) used by GTDB[35,36] 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[59] 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.

SNP annotation and gene enrichment analysis

Synonymous and non-synonymous SNPs were annotated using SnpEff v4.3t[60], 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[61].

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[56].

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 on Zenodo (DOI: 10.5281/zenodo.14880462) and Supplementary Tables. Source data are provided with this paper.

References

1. Charlesworth, B. & Charlesworth, D. Population genetics from 1966 to 2016. *Heredity* **118** (2017).
2. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17** (2016).
3. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nature Reviews Genetics* **22** (2021).
4. Blanchet, S., Prunier, J. G. & De Kort, H. Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends in Genetics* **33**, 579–580 (2017).
5. Leigh, D. M. *et al.* Opportunities and challenges of macrogenetic studies. *Nature Reviews Genetics* **22**, 791–807 (2021).
6. Manel, S. *et al.* Global determinants of freshwater and marine fish genetic diversity. *Nat Commun* **11** (2020).
7. Theodoridis, S. *et al.* Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nat Commun* **11** (2020).
8. Vellend, M. & Geber, M. A. Connections between species diversity and genetic diversity. *Ecology Letters* **8** (2005).
9. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Nature Reviews Genetics* **17**, 422–433 (2016).
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. *Proceedings of the National Academy of Sciences* **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. Reimer, L. C. *et al.* BacDive in 2022: The knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res* **50**, (2022).
16. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* **7**, (2020).
17. 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).
18. Lynch, M. & Conery, J. S. The Origins of Genome Complexity. *Science (1979)* **302**, 1401–1404 (2003).
19. Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc Natl Acad Sci U S A* **113**, 11399–11407 (2016).
20. Bourguignon, T. *et al.* Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. *Current Biology* **30**, 3848-3855.e4 (2020).
21. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* **6** (2010).
22. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9** (2008).
23. Lin, M. & Kussell, E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods* **16**, 199–204 (2019).
24. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput Biol* **11**, 1–18 (2015).
25. Bobay, L. M., Ellis, B. S. H. & Ochman, H. ConSpeciFix: Classifying prokaryotic species based on gene flow. *Bioinformatics* **34**, (2018).
26. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17** (2016).
27. Dewar, A. E., Hao, C., Belcher, L. J., Ghoul, M. & West, S. A. Bacterial lifestyle shapes pangenomes. *Proceedings of the National Academy of Sciences* **121** (2024).
28. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME Journal* **11**, 1719–1721 (2017).
29. Bobay, L.-M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol Biol* **18**, 153 (2018).
30. 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).
31. Dey, A., Chan, C. K. W., Thomas, C. G. & Cutter, A. D. Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri. *Proc Natl Acad Sci U S A* **110**, (2013).
32. Yang, C., Cui, Y., Didelot, X., Yang, R. & Falush, D. Why panmictic bacteria are rare. *bioRxiv* (2018).
33. Cui, Y. *et al.* The landscape of coadaptation in Vibrio parahaemolyticus. *Elife* **9**, e54136 (2020).
34. Tourrette, E. *et al.* An ancient ecospecies of Helicobacter pylori. *Nature* **635**, 178–185 (2024).
35. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996 (2018).

36. 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).
37. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, (2001).
38. 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).
39. Liu, S., Svensson, S. L. & Falush, D. A putative ecocline in Klebsiella pneumoniae. *bioRxiv* (2025).
40. 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).
41. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to Identify Similar Regions in Large Sequence Sets . *Curr Protoc Bioinformatics* **00**, (2003).
42. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016).
43. 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).
44. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
45. 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).
46. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
47. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
48. 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).
49. Yang, C. *et al.* Recent mixing of Vibrio parahaemolyticus populations. *ISME J* **13**, 2578–2588 (2019).
50. Mitrofanov, A. *et al.* CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res* **49**, (2021).
51. 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).
52. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**, (2017).
53. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, (2007).
54. Barrett, J. C. Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc* **4**, (2009).
55. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).

56. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).
57. 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).
58. Tung Ho, L. S. & Ané, C. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst Biol* **63**, (2014).
59. 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).
60. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, (2012).
61. 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., H.H. and N.W. contributed to data collectionand and performed bioinformatics analysis. 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.

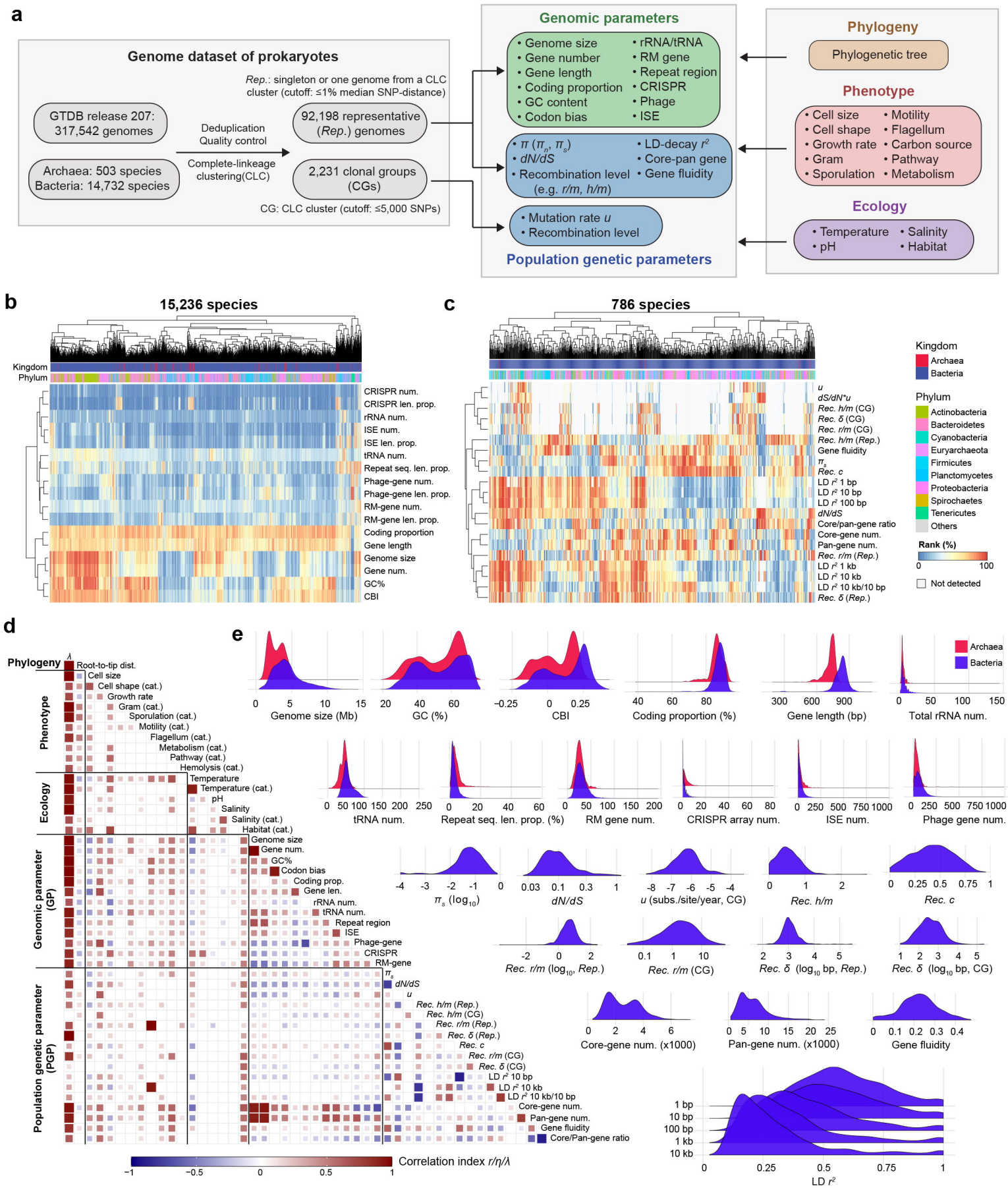
Figure legends

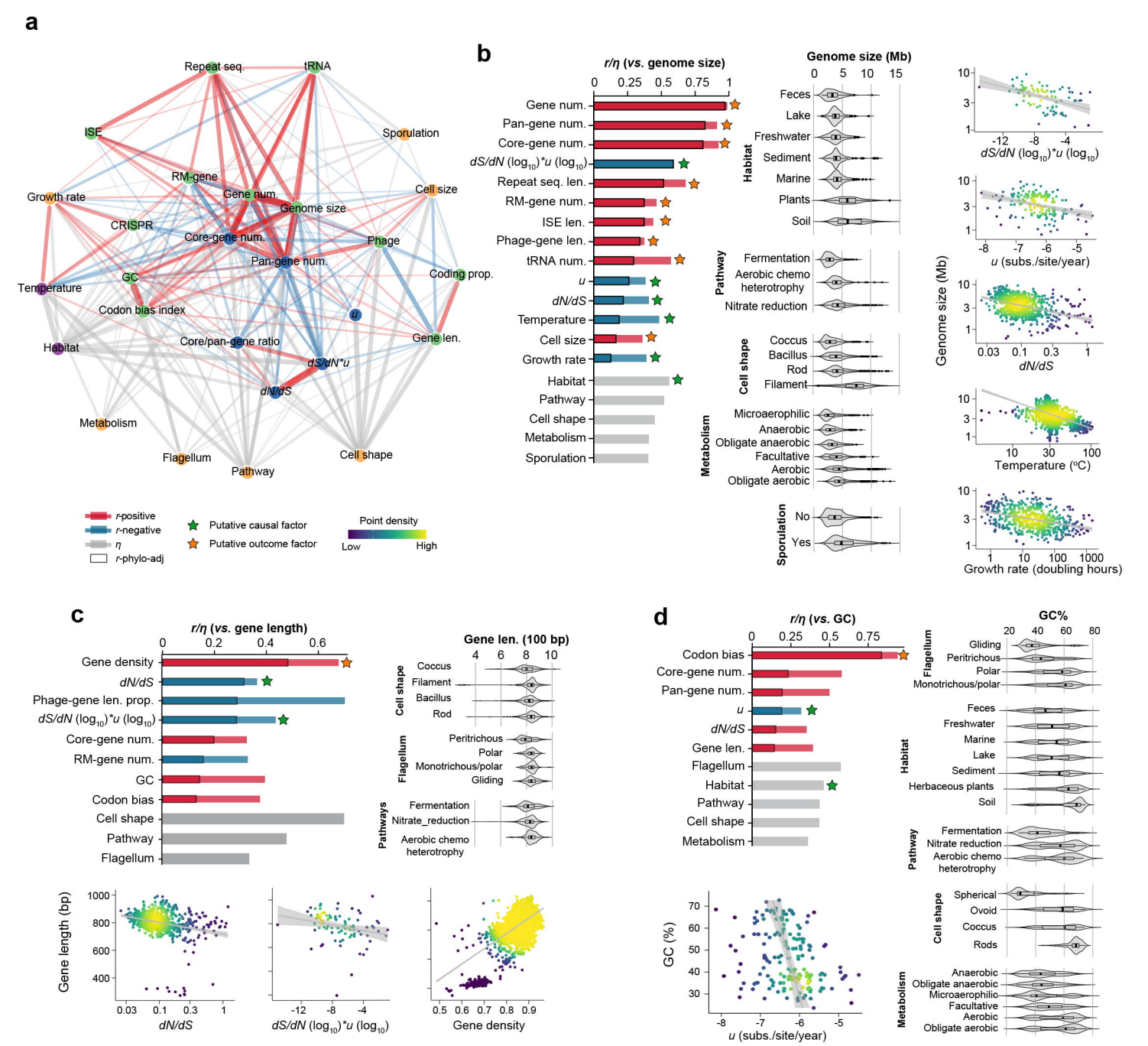
Figure 1. Overview of the macrogenetic atlas of prokaryotes (MAP). **a)** Study flowchart. **b,c)** Heatmaps displaying the rank percentages of genomic parameters (GP) and population genetic parameters (PGP) for each species. The bars at the top indicate the kingdom and phylum of each species, while the heatmap colors represent rank percentages. **d)** Correlation heatmap among various parameters. Phylogenetic signals were quantified using Pagel's λ. **e)** Distribution of different GP and PGP. Red and blue colors represent archaea and bacteria, respectively. PGP is shown only for bacteria, as few archaeal species met the criteria for PGP estimation (≥10 *Rep.* genomes).

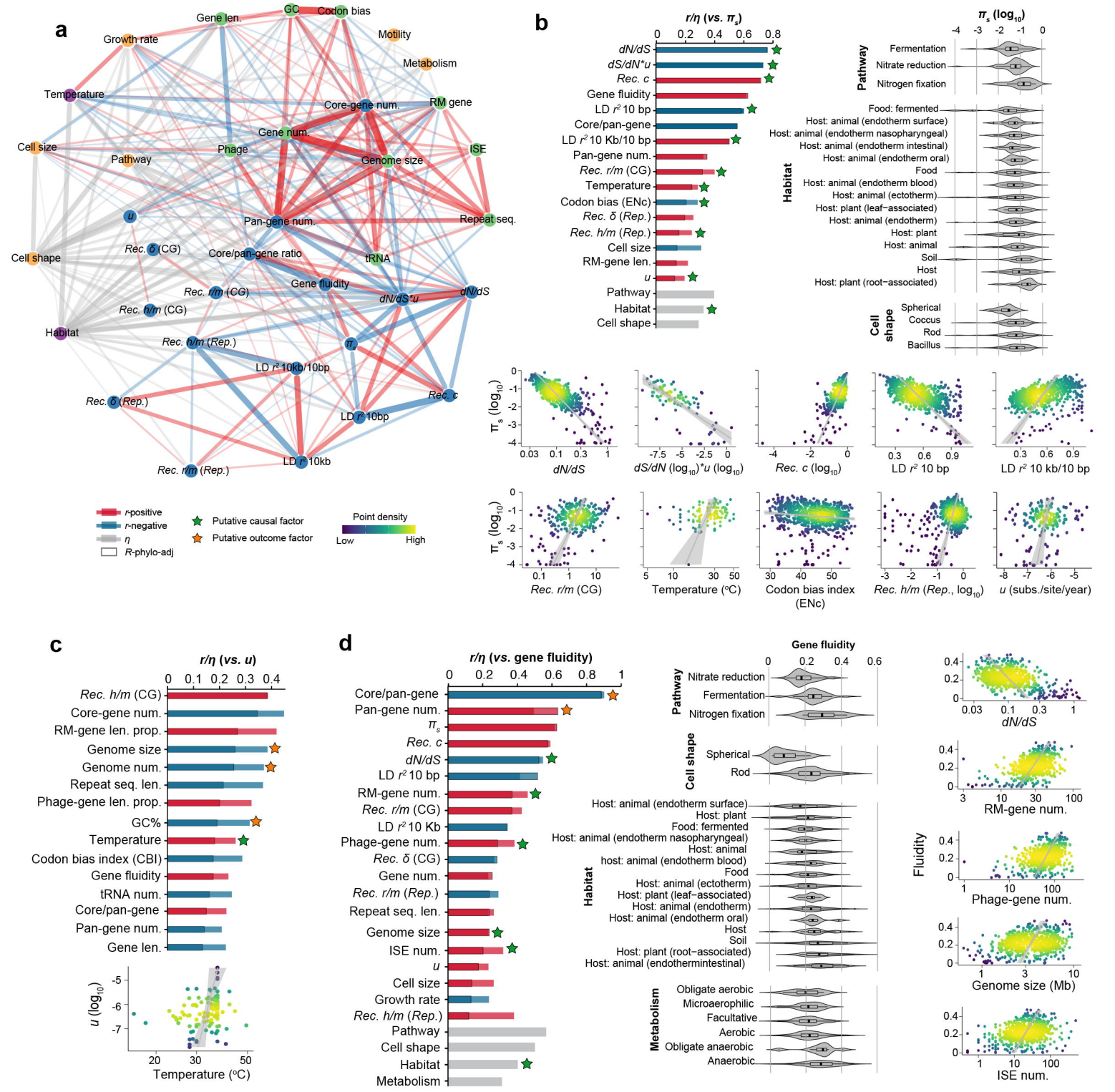
Figure 2. Determinants of genomic characteristics. a) Correlation network of three GPs (genome size, gene length and GC content, r or $\eta \geq 0.3$) and related parameters. Node color represents parameter categories, while line color indicates the type or direction of the correlation. Line width and transparency indicate the strength of the correlation coefficient. **b-d)** Three GPs related parameters before and after phylogenetic signal adjustment ($r\text{-}adj \geq 0.1$). Bar color indicates the direction or type of correlation. Green and orange stars indicate putative causal and consequential factors, respectively. Violin and box plots display the distribution of GPs across different categories, with only categories containing ≥100 species shown. Point density plots show detailed correlations between the GPs and their putative causal/outcome factors, with color representing point density.

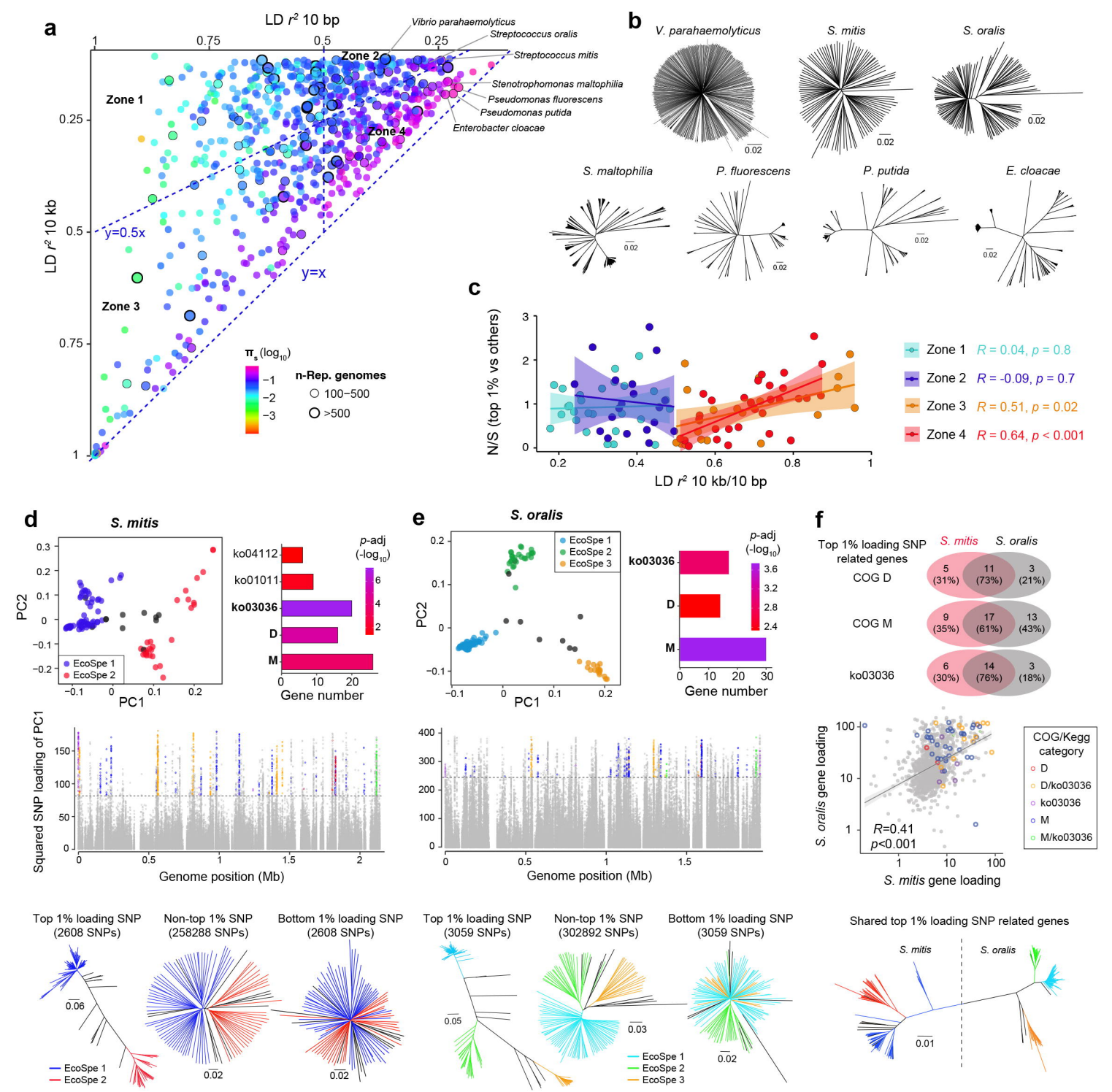
Figure 3. Determinants of population genetics characteristics. a) Correlation network of three PGs (genetic diversity, mutation rate, and gene fluidity, r or $\eta \geq 0.2$) and their related parameters. Node and line color, width and transparency are as in Fig. 2a. **b-d)** Three PGs related parameters before and after phylogenetic signal adjustment ($r\text{-}adj \geq 0.1$). Bar and point colors are as in Fig. 2b-d. Only categories containing ≥10 species are shown.

Figure 4. Determinants of population structure and ecospecies. a) Relationships between short-range LD (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)** Phylogenetic trees of seven selected species. **c)** Correlations between LD saturation and N/S (top 1% vs other SNPs), 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 the ratio of N/S for SNPs outside the top 1% (SNP-other-PC1). **d,e)** Top: Principal components analysis (left, point colors indicate inferred ecospecies) and functional enrichment of genes related to SNP-top1%-PC1 (right) in *Streptococcus mitis* and *S. oralis.* Middle: Distribution of squared SNP loadings on PC1 across the genome, with point colors indicating COG or KEGG categories. Dashed lines represent the top 1% SNP loading cutoff. Bottom: Phylogenetic trees based on SNPs with top and bottom 1% SNP loadings on PC1 and other SNPs. Branch colors indicate inferred ecospecies. **f)** Top: Shared homologous genes associated with SNP-top1%-PC1 between two species. Middle: Correlation of gene loading (average SNP loading per gene) between species. Point colors indicate COG or KEGG categories. Bottom: Phylogenetic trees based on protein sequences of shared genes associated with SNP-top1%-PC1 between two 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 • Gene number • Gene length • Coding proportion • GC content • Codon bias
• rRNA/tRNA • RM gene • Repeat region • CRISPR • Phage • ISE
• π (πn, πs) • dN/dS • Recombination level (e.g. r/m, h/m)
• LD-decay r2 • Core-pan gene • Gene fluidity
• Mutation rate u • Recombination level
Population genetic parameters
Phylogeny
Phylogenetic tree
Phenotype
• Cell size • Cell shape • Growth rate • Gram • Sporulation
• Motility • Flagellum • Carbon source • Pathway • Metabolism
Ecology
• Temperature • pH • Salinity • Habitat
b
15,236 species
Kingdom
Phylum
CRISPR num.
CRISPR len. prop.
rRNA num.
ISE num.
ISE len. prop.
tRNA num.
Repeat seq. len. prop.
Phage-gene num.
Phage-gene len. prop.
RM-gene num.
RM-gene len. prop.
Coding proportion
Gene length
Genome size
Gene num.
GC%
CBI
c
786 species
u
dS/dN*u
Rec. h/m (CG)
Rec. δ (CG)
Rec. r/m (CG)
Rec. h/m (Rep.)
Gene fluidity
πs
Rec. c
LD r2 1 bp
LD r2 10 bp
LD r2 100 bp
dN/dS
Core/pan-gene ratio
Core-gene num.
Pan-gene num.
Rec. r/m (Rep.)
LD r2 1 kb
LD r2 10 kb
LD r2 10 kb/10 bp
Rec. δ (Rep.)
Kingdom
Archaea
Bacteria
Phylum
Actinobacteria
Bacteroidetes
Cyanobacteria
Euryarchaeota
Firmicutes
Planctomycetes
Proteobacteria
Spirochaetes
Tenericutes
Others
Rank (%)
0
100
Not detected
d
Phylogeny
λ
Root-to-tip dist.
Phenotype
Cell size
Cell shape (cat.)
Growth rate
Gram (cat.)
Sporulation (cat.)
Motility (cat.)
Flagellum (cat.)
Metabolism (cat.)
Pathway (cat.)
Hemolysis (cat.)
Ecology
Temperature
Temperature (cat.)
pH
Salinity
Salinity (cat.)
Habitat (cat.)
Genomic parameter (GP)
Genome size
Gene num.
GC%
Codon bias
Coding prop.
Gene len.
rRNA num.
tRNA num.
Repeat region
ISE
Phage-gene
CRISPR
RM-gene
Population genetic parameter (PGP)
πs
dN/dS
u
Rec. h/m (Rep.)
Rec. h/m (CG)
Rec. r/m (Rep.)
Rec. δ (Rep.)
Rec. c
Rec. r/m (CG)
Rec. δ (CG)
LD r2 10 bp
LD r2 10 kb
LD r2 10 kb/10 bp
Core-gene num.
Pan-gene num.
Gene fluidity
Core/Pan-gene ratio
−1
−0.5
0
0.5
1
Correlation index r/η/λ
e
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.
ISE num.
Phage gene num.
πs (log10)
dN/dS
u (subs./site/year, CG)
Rec. h/m
Rec. c
Rec. r/m (log10, 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
Repeat seq.
tRNA
ISE
Sporulation
RM-gene
Growth rate
Gene num.
Cell size
Genome size
CRISPR
Core-gene num.
Phage
Pan-gene num.
GC
Coding prop.
Temperature
Codon bias index
u
Gene len.
Habitat
Core/pan-gene ratio
dS/dN*u
dN/dS
Metabolism
Flagellum
Cell shape
Pathway
r-positive
r-negative
η
r-phylo-adj
Putative causal factor
Putative outcome factor
Point density
Low
High
b
r/η (vs. genome size)
0 0.25 0.5 0.75 1
Gene num.
Pan-gene num.
Core-gene num.
dS/dN (log10)*u (log10)
Repeat seq. len.
RM-gene num.
ISE len.
Phage-gene len.
tRNA num.
u
dN/dS
Temperature
Cell size
Growth rate
Habitat
Pathway
Cell shape
Metabolism
Sporulation
Genome size (Mb)
0 5 10 15
Habitat
Feces
Lake
Freshwater
Sediment
Marine
Plants
Soil
Pathway
Fermentation
Aerobic chemo heterotrophy
Nitrate reduction
Cell shape
Coccus
Bacillus
Rod
Filament
Metabolism
Microaerophilic
Anaerobic
Obligate anaerobic
Facultative
Aerobic
Obligate aerobic
Sporulation
No
Yes
dS/dN (log10)*u (log10)
u (subs./site/year)
Genome size (Mb)
dN/dS
Temperature (°C)
Growth rate (doubling hours)
c
r/η (vs. gene length)
0 0.2 0.4 0.6
Gene density
dN/dS
Phage-gene len. prop.
dS/dN (log10)*u (log10)
Core-gene num.
RM-gene num.
GC
Codon bias
Cell shape
Pathway
Flagellum
Gene len. (100 bp)
4 6 8 10
Cell shape
Coccus
Filament
Bacillus
Rod
Flagellum
Peritrichous
Polar
Monotrichous/polar
Gliding
Pathways
Fermentation
Nitrate_reduction
Aerobic chemo heterotrophy
Gene length (bp)
dN/dS
dS/dN (log10)*u (log10)
Gene density
d
r/η (vs. GC)
0 0.25 0.5 0.75
Codon bias
Core-gene num.
Pan-gene num.
u
dN/dS
Gene len.
Flagellum
Habitat
Pathway
Cell shape
Metabolism
GC%
20 40 60 80
Flagellum
Gliding
Peritrichous
Polar
Monotrichous/polar
Habitat
Feces
Freshwater
Marine
Lake
Sediment
Herbaceous plants
Soil
Pathway
Fermentation
Nitrate reduction
Aerobic chemo heterotrophy
Cell shape
Spherical
Ovoid
Coccus
Rods
Metabolism
Anaerobic
Obligate anaerobic
Microaerophilic
Facultative
Aerobic
Obligate aerobic
GC (%)
u (subs./site/year)


a
GC
Codon bias
Gene len.
Growth rate
Motility
Metabolism
Temperature
Core-gene num.
RM gene
Gene num.
Cell size
Phage
ISE
Pathway
Genome size
u
Pan-gene num.
Repeat seq.
Rec. δ (CG)
Core/pan-gene ratio
Cell shape
tRNA
Gene fluidity
Rec. r/m (CG)
dN/dS*u
dN/dS
Rec. h/m (CG)
Habitat
Rec. h/m (Rep.)
π_s
LD r² 10kb/10bp
Rec. δ (Rep.)
Rec. c
LD r² 10bp
LD r² 10kb
Rec. r/m (Rep.)
r-positive
r-negative
η
R-phylo-adj
Putative causal factor
Putative outcome factor
Point density
Low
High
b
r/η (vs. π_s)
dN/dS
dS/dN*u
Rec. c
Gene fluidity
LD r² 10 bp
Core/pan-gene
LD r² 10 Kb/10 bp
Pan-gene num.
Rec. r/m (CG)
Temperature
Codon bias (ENc)
Rec. δ (Rep.)
Rec. h/m (Rep.)
Cell size
RM-gene len.
u
Pathway
Habitat
Cell shape
π_s (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)
Cell shape
Spherical
Coccus
Rod
Bacillus
π_s (log10)
dN/dS
dS/dN (log10)*u (log10)
Rec. c (log10)
LD r² 10 bp
LD r² 10 kb/10 bp
Rec. r/m (CG)
Temperature (°C)
Codon bias index (ENc)
Rec. h/m (Rep., log10)
u (subs./site/year)
c
r/η (vs. u)
Rec. h/m (CG)
Core-gene num.
RM-gene len. prop.
Genome size
Genome num.
Repeat seq. len.
Phage-gene len. prop.
GC%
Temperature
Codon bias index (CBI)
Gene fluidity
tRNA num.
Core/pan-gene
Pan-gene num.
Gene len.
u (log10)
Temperature (°C)
d
r/η (vs. gene fluidity)
Core/pan-gene
Pan-gene num.
π_s
Rec. c
dN/dS
LD r² 10 bp
RM-gene num.
Rec. r/m (CG)
LD r² 10 Kb
Phage-gene num.
Rec. δ (CG)
Gene num.
Rec. r/m (Rep.)
Repeat seq. len.
Genome size
ISE num.
u
Cell size
Growth rate
Rec. h/m (Rep.)
Pathway
Cell shape
Habitat
Metabolism
Gene fluidity
Pathway
Nitrate reduction
Fermentation
Nitrogen fixation
Cell shape
Spherical
Rod
Habitat
Host: animal (endotherm surface)
Host: plant
Food: fermented
Host: animal (endotherm nasopharyngeal)
Host: animal
host: animal (endotherm blood)
Food
Host: animal (ectotherm)
Host: plant (leaf-associated)
Host: animal (endotherm)
Host: animal (endotherm oral)
Host
Soil
Host: plant (root-associated)
Host: animal (endothermintestinal)
Metabolism
Obligate aerobic
Microaerophilic
Facultative
Aerobic
Obligate anaerobic
Anaerobic
dN/dS
RM-gene num.
Fluidity
Phage-gene num.
Genome size (Mb)
ISE num.


a
LD r² 10 bp
LD r² 10 kb
Zone 1
Zone 2
Zone 3
Zone 4
y=0.5x
y=x
Vibrio parahaemolyticus
Streptococcus oralis
Streptococcus mitis
Stenotrophomonas maltophilia
Pseudomonas fluorescens
Pseudomonas putida
Enterobacter cloacae
πs (log10)
n-Rep. genomes
100–500
>500
b
V. parahaemolyticus
S. mitis
S. oralis
S. maltophilia
P. fluorescens
P. putida
E. cloacae
c
N/S (top 1% vs others)
LD r² 10 kb/10 bp
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
d
S. mitis
EcoSpe 1
EcoSpe 2
PC1
PC2
ko04112
ko01011
ko03036
D
M
p-adj (-log10)
Gene number
Squared SNP loading of PC1
Genome position (Mb)
e
S. oralis
EcoSpe 1
EcoSpe 2
EcoSpe 3
PC1
PC2
ko03036
D
M
p-adj (-log10)
Gene number
Genome position (Mb)
f
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 SNP (2608 SNPs)
Non-top 1% SNP (258288 SNPs)
Bottom 1% loading SNP (2608 SNPs)
Top 1% loading SNP (3059 SNPs)
Non-top 1% SNP (302892 SNPs)
Bottom 1% loading SNP (3059 SNPs)
Shared top 1% loading SNP related genes
S. mitis
S. oralis
EcoSpe 1
EcoSpe 2
EcoSpe 3